Journal of Information and Data Management, 2023, 14:1, doi: 10.5753/jidm.2023.3181

Using Active Learning for Segmentation and Semantic Classification of Legal Acts Extracted from Official Diaries

Kattiana Constantino [**Universidade Federal de Minas Gerais** | ***kattiana@dcc.ufmg.br***]
Thiago H. P. Silva [**Universidade Tecnológica Federal do Paraná** | ***thiagosilva@utfpr.edu.br***]
João Vítor B. Silva [**Universidade Federal de Minas Gerais** | ***joao.silva@dcc.ufmg.br***]
Victor Augusto L. Cruz [**Universidade Federal de Minas Gerais** | ***victoraugusto@dcc.ufmg.br***]
Otávio M. M. Zucheratto [**Universidade Federal de Minas Gerais** | ***otaviozucheratto@dcc.ufmg.br***]
Marcos Carvalho [**Universidade Federal de Minas Gerais** | ***marcoscarvalho@dcc.ufmg.br***]
Welton Santos [**Universidade Federal de Minas Gerais** | ***weltonsantos@dcc.ufmg.br***]
Celso França [**Universidade Federal de Minas Gerais** | ***celsofranca@dcc.ufmg.br***]
Claudio M. V. de Andrade [**Universidade Federal de Minas Gerais** | ***claudio.valiense@dcc.ufmg.br***]
Alberto H. F. Laender [**Universidade Federal de Minas Gerais** | ***laender@dcc.ufmg.br***]
Marcos André Gonçalves [**Universidade Federal de Minas Gerais** | ***mgoncalv@dcc.ufmg.br***]

Department of Computer Science, Universidade Federal de Minas Gerais, 31270-901, Belo Horizonte, MG, Brazil.

Received: 27 February 2023 • **Published:** 20 October 2023

Abstract. Based on openness and transparency for good governance, unimpeded and verifiable access to legal and regulatory information is essential. With such access, we can monitor government actions to ensure that public financial resources are not improperly or inconsistently used. This facilitates, for example, the detection of unlawful behavior in public actions, such as bidding processes and auctions. However, different public agencies have their own criteria for standardizing the models and formats used to make information available, as exemplified in the varying styles observed in municipal, state, and union (federal) documents. In this context, we aim to minimize the effort to deal with public documents, notably official gazettes. For this, we propose a structure-oriented heuristic for extracting relevant excerpts from their texts. We then characterize these excerpts through morphosyntactic analysis and entity recognition. Subsequently, we semantically classify the extracted fragments into "sections of interest" (e.g., bids, laws, personnel, budget) using an active learning strategy to reduce the manual labeling effort. We also improve the classification process by incorporating transformers, stacking, and by combining different types of representations (e.g., frequentist, static, and contextual semantic embeddings). Furthermore, we exploit oversampling based on semi-supervised learning to deal with (labeled) data scarceness and skewness. Finally, we combine all these contributions in a real-time annotation tool with active learning support that achieves 100% accuracy in extraction and an overall accuracy of 85% in classification with very little labeling effort.

Keywords: Semantic classification, Active Learning, Official Diaries, Annotation tool

1 Introduction

Access to public data is relevant not only to observe the decisions of the federated states[1], but also to monitor how public policies aimed at the population are defined and implemented, thus making it possible to democratize bidding processes and public auctions, as well as to monitor the expenses of each government agency or institution. In this context, it is also possible to monitor how public policies aimed at the population are defined and executed. For example, with more transparency in bids or auctions, we can detect fraud in the revenues and expenses of each federal state [Pinto *et al.*, 2021, 2023; Rangel *et al.*, 2020; Silva *et al.*, 2023].

In Brazil, the *Access to the Information Law*[2], of its 1988 Federal Constitution, ensures full access to public data. With this law, it became mandatory to make public data available on the official websites of each one of the federated states. In this way, it is possible to propose several analyzes with the aid of statistical and computational techniques to monitor government activities, such as, for example, the detection of fraud in bidding processes.

However, the problem is the lack of legal acts standardization, such as different document format types. For instance, the federated entities adopt different forms of document elaboration without the support of a precise official management model. Furthermore, there are no rules on how documents must be published on their respective websites, or even labels associated with them for better information retrieval.

This article addresses an effort to deal with two essential aspects of official government diaries: *text segmentation* and *semantic classification*. The former consists of separating the text of an official diary into sections. In this way, it allows us to identify the associated federal entities involved, the title and the content of the published governmental acts, and the person(s) responsible for their creation. This is not a trivial problem, as each document has a specific presentation

[1] A federated state is a territorial and constitutional community forming part of a federation.

[2] Law No. 12,527 of November 18, 2011, accessed 22 June 2023, `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`.

structure with different graphical components (e.g., visual separators). Regarding the classification task, the aim is to correctly predict to which class an extracted textual excerpt belongs, given a predetermined set of categories such as *Law*, *Bid*, *Budget* or *Personnel*. In order to address this problem, we propose an active learning strategy that allows repeated interactions between the classifier and the annotator, selecting a reduced number of the most valuable instances for training [de Freitas *et al*., 2010; Campos *et al*., 2017b; **?**].

In our preliminary work [Constantino *et al*., 2022], we presented two key contributions that address the aforementioned problems: we first introduced a heuristic-based strategy that leverages structural aspects to extract text segments from official diaries, and then we proposed a semantic strategy that utilizes active machine learning and state-of-the-art transformer classifiers to classify these extracted segments. Building upon our previous findings, this article expands our research by providing the following additional contributions: (i) a more in-depth characterization of our dataset, focusing specifically on its semantic aspects; (ii) a comprehensive discussion on the development of a proposed annotation tool, which works as a Web Service as part of the classification process; and (iii) extensive experimentation and analysis involving advanced classification strategies, including stacking [Gomes *et al*., 2021] and the combination of diverse text representations [de Andrade, Claudio Moisés Valiense and Gonçalves, Marcos André, 2020]. These strategies aim to enhance the classification results and yield further insights. By elaborating on these aspects, our article offers a more comprehensive and detailed perspective on our research, extending and enhancing our preliminary work.

Initially, our work was subdivided into two fronts, addressing the tasks of (1) *text segmentation* and (2) *semantic classification*. The segmentation task is the process of separating a text into useful blocks, such as sentences, paragraphs, or sections [Fernandes *et al*., 2007; Pak and Teh, 2018]. As illustrated in Figure 1, in the context of an official diary, our interest is to extract excerpts containing the name of a government entity followed by the title/subtitle, body and signatures of a specific document issued by that specific entity (ex., a municipal decree). Considering that most scenarios in public administration involve the publication of such documents in PDF format without annotations, in our previous work [Constantino *et al*., 2022] we proposed a heuristic that explores structural aspects for extracting excerpts from such documents. Specifically, such strategy explores the distance between textual elements to identify the margins that delimit all the contents belonging to it, from lists of characters forming words to a set of lines forming a block. As a result, the segments had their textual content duly extracted (accuracy of 100%).

In this work, besides further characterizing and analyzing our segmentation solutions, we considerably extend the second step focused on the classification task. The classification task seeks to predict which class or category the extracted excerpt belongs to, given a set of pre-existing categories [Cunha *et al*., 2021]. For example, if part of an extracted section includes the text fragment *"... RESOLVE: ART.1º - EXONERAR a servidora ..."*[3], its class corresponds to *Human Resources*, while a subtitle of an excerpt beginning with *"... Regula o acesso a informações previsto no inciso ..."*[4] can be considered as belonging to the class of *Laws*.

More specifically, the classification task takes as input a set of government acts and, for each one, defines as output the semantic class it belongs to, considering the following options: (i) Bids; (ii) Law; (iii) Budget; and (iv) Personnel (e.g., nominations and other Human resources-related acts). In this work, we aim to determine which are the best machine learning techniques to deal with the correct assignment of the aforementioned semantic classes. As we will see in the next sections, the most suitable solution consists of exploiting transformers – a deep learning model that adopts the mechanism of self-attention – together with an active learning technique – a strategy that selects the best samples to be labeled by experts. Indeed, we tested several advanced classification techniques found in the literature, such as Stacking [Gomes *et al*., 2021] and a combination of representations [de Andrade, Claudio Moisés Valiense and Gonçalves, Marcos André, 2020]. Furthermore, we exploit oversampling based on semi-supervised learning to deal with (labeled) data scarceness and skewness. Despite all these efforts, we could not obtain statistically superior results, showing that our current solutions [Constantino *et al*., 2022] have probably achieved the limits of what can be done by using the current state-of-the-art solutions for automated text classification. In any case, our final recommendations suggest using Stacking, even though it does not achieve a good tradeoff between per-class effectiveness and cost.

This work represents a collaborative effort between the Department of Computer Science at the Federal University of Minas Gerais and the Public Ministry of the State of Minas Gerais (MPMG)[5]. The primary objective of this partnership is to conduct in-depth analyses of extensive public data repositories to characterize public expenditures, thereby providing crucial support for complex investigations.

Specifically, our analyses have made significant contributions to data management and classification solutions, focusing on textual information extraction and semantic classification. From a technological standpoint, the outcome of this work is the development and dissemination of open-source tools, thereby democratizing access to efficient analysis of public data.

It is worth noting that the scope of this work, as determined by the MPMG, pertains to the classification of public data. The MPMG has selected priority cases based on its internal procedures. We have limited our analysis to a repository comprising a generic collection encompassing various federated entities to avoid favoritism towards specific instances. Consequently, we were provided with the official repository for the year 2020, containing all Official Diaries from the Association of the Municipalities of the State of Minas Gerais, Brazil[6]. This repository consists of 1,640 documents in PDF

[3] A free English translation of this Portuguese text excerpt is: *"... RESOLUTION: 1st Article - EXONERATE the government employee ..."*

[4] A free English translation of this Portuguese text excerpt is: *"... Regulates access to information provided for item ..."*

[5] Public Ministry of the State of Minas Gerais, accessed 22 June 2023, `https://www.mpmg.mp.br`.

[6] Association of Municipalities of the State of Minas Gerais, accessed

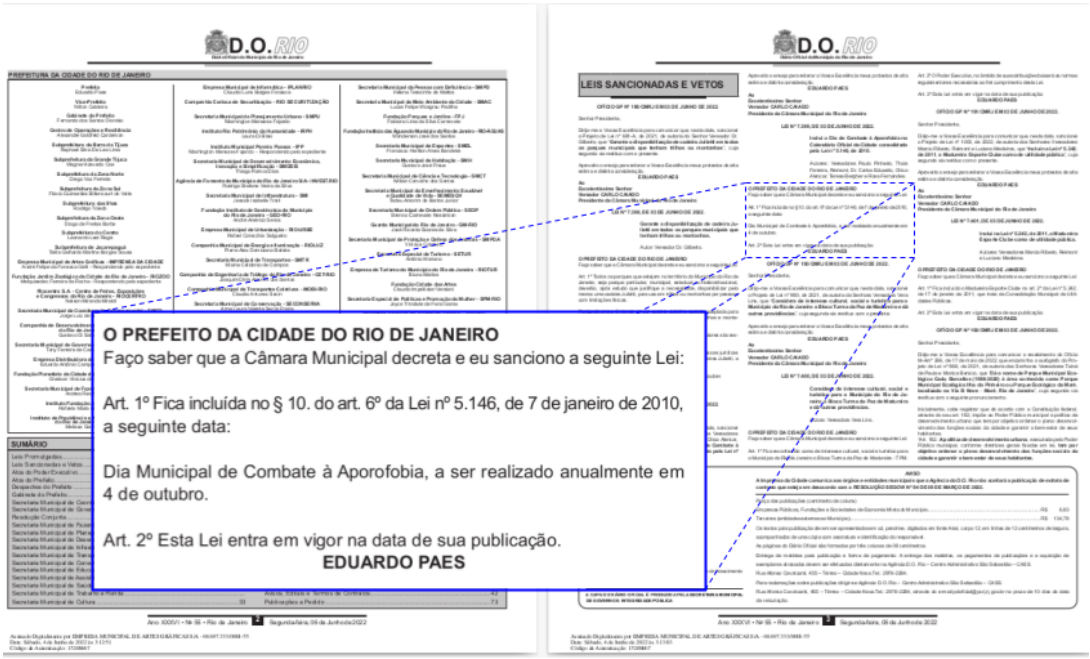

Figure 1. Example of the content from an edition of the Official Diary of the City of Rio de Janeiro, showcasing multiple administrative acts issued by several sectors of the city administration. Notably, the highlighted act announces a specific official date to combat aporophobia, decreed by the city's mayor, Eduardo Paes.

format, affiliated with hundreds of entities, resulting in a diverse range of excerpts that present an interesting scenario for evaluating our semantic classification process.

The rest of this article is organized as follows. Section 2 addresses related work. Then, Section 3 provides a characterization of the addressed official diaries, whereas Section 4 discusses the segmentation process applied to such diaries and Section 5 presents the semantic classification applied to them. Then, Section 6 describes our experimental setup and Section 7 summarizes the main results obtained. Finally, Section 8 summarizes our main conclusions and provides some insights for future work.

2 Related Work

This section presents a brief overview of some previous works whose goals are similar to ours, but follow distinct approaches. Text mining is incredibly challenging for some domains, such as legal and legislative texts. They often contain ambiguous and vague legal terms and are typically context- and time-dependent [Knackstedt *et al.*, 2014]. For this reason, the automatic processing of legal/official documents has been extensively researched. In the context of the problem of text segmentation and semantic classification of information involving data from public administration, some works have emphasized that the steps of treatment and availability are considered as prerequisites for the analysis of more complex data.

For instance, Lin *et al.* [2015] proposed a text mining mechanism to automatically classify legislative documents that refer to each legislator, and then represent the proportion of their legislative performance on certain categories. They used the SVM method to build a model to classify a new document to the appropriate category. In order to maintain the classification categories up to dated, in their work they also evaluated the difference between labeling contents by domain experts and by the general public. In our context, we collected official government diaries and classified their segmented texts by applying the following techniques: Support Vector Machines (SVM), Bidirectional Encoder Representations from Transformers (BERT), BERT model for Brazilian Portuguese (BERTimbau), Language-agnostic BERT Sentence Embedding (LaBSE) model, FastText Concat (TF-IDF, FastText), Concat (TF-IDF, BERTimbau) and Stacking.

Rangel *et al.* [2020] applied supervised machine learning techniques to infer the categories (e.g., *health* or *finance*) of documents available in government data portals, while Pinto *et al.* [2021] and **?** performed textual extractions from official diaries by using regular expressions and built, as a result, a knowledge base according to a grammar defined specifically for issuing acts for moving personnel of Rio de Janeiro City Hall. In our work, we characterized a large sample of official diaries morphosyntactic analyses and entity recognition.

Regarding the problem of lack of standardization for designating public service categories, Pereira *et al.* [2021] addressed it by proposing a taxonomy to classify better the Brazilian government data involved. Also, it is worth mentioning a tool for document annotation and classification proposed by Inuzuka *et al.* [2020] in partnership with a private company. They used an active learning technique to classify whether the information contained in an excerpt from the Official Gazette is or is not of legal content. In our work, we also adopted classification techniques and details the proposal of an annotation tool integrated into the classification process with active learning to identify the major semantic topics in our sample of official diaries.

22 June 2023, `https://www.diariomunicipal.com.br/amm-mg/`.

Table 1. Functions of terms present in the sentences.

Part-of-speech	Average	Median	SD
Proper Noun	28.8%	27.6%	8.6%
Punctuation	17.0%	17.6%	2.4%
Noun	15.8%	15.5%	2.9%
Numeral	10.1%	9.4%	2.8%
Preposition	9.4%	9.7%	1.3%
Determiner	7.3%	7.6%	1.2%
Verb	3.6%	3.6%	0.8%
Adjective	3.2%	3.2%	0.8%
Coord. Conjunction	1.9%	1.8%	0.7%
Auxiliary Verb	0.5%	0.4%	0.2%
Pronoun	0.5%	0.5%	0.1%
Others	0.4%	0.1%	0.4%

3 Characterization of Official Diaries Excerpts

In this section, we present details about morphosyntactic aspects of the dataset. We emphasize that such characterization does not integrate the segmentation and classification stages. In contrast, this characterization aims to describe and understand the particularities observed in the set of governmental acts contained in official diaries, making it possible to obtain information and patterns that might help advance future proposals of semantic classification. Observing the content of the document sections, we identify specific terms that can be good indicators to classify the semantics emitted in each publication. For example, one of the ways of publishing the professional admission to a public office has the following style: "FILL", under the terms of [item/law], the public servant: [full name of the admitted]"[7]. That is, when the verb "fill" appears in *the third person singular in the present indicative* in a specific segment, it is likely to be classified as the class "admission act". Also, note that in this fragment, there is a standardization of the verb in capital letters. Thus, we can reinforce that specific terms in an act can be helpful indicators to determine the semantics of a publication.

Next, we show the results of the characterization of a large sample of official diaries that includes a total of 416 PDF documents corresponding to all issues of an entire year. This sample, after being converted into a textual format, resulted in a single textual document of 342.7 MB, with 5.8 million lines and 49.5 million words. We then divided the characterization of this document into two parts: morphosyntactic analysis (Subsection 3.1) and entity recognition (Subsection 3.2).

3.1 Morphosyntactic Analysis

In this subsection, we use natural language processing techniques to characterize official dairies. In particular, we aim to determine the patterns of terms concerning the functions they play in a sentence (syntactic analysis) and the behavior of such terms independently of their connection with other words (morphological analysis).

[7] A loose translation of the following Portuguese fragment: *"LOTA, nos termos do [inciso/lei], o(a) servidor(a): [nome completo do admitido]"*.

Table 2. Nominal forms of verbs present in the sentences.

Form	Average	Median	SD
Finite	39.7%	39.6%	4.7%
Participle	39.2%	38.6%	5.4%
Infinitive	15.5%	15.0%	3.3%
Gerund	5.6%	5.4%	1.3%

Table 3. Verb tenses present in the sentences.

Tense	Average	Median	SD
Present	27.9%	27.6%	4.1%
Past	4.5%	4.0%	2.5%
Future	5.9%	5.7%	2.0%
Imperfect	0.9%	0.8%	0.8%
Pluperfect	0.0%	0.0%	0.1%
Not detected	60.8%	60.8%	4.7%

Regarding the syntactic analysis, Table 1 shows the proportion of syntactic functions of terms present in the sentences. For example, on average, each document has about 29% of proper nouns. Note that the means and medians are very close, indicating that the documents tend to have common characteristics. Furthermore, the standard deviation values (except for proper names) are low.

Since verbs are very relevant syntactic words for the problem at hand, they were characterized as they indicate actions, states or events. First, we report that there are 177 different variations of the morphological characteristics present in the repository. In particular, it is noteworthy that the verbs are mainly in the third person singular. Tables 2, 3, and 4 highlight well-defined patterns with very close means and medians in all official diaries, as detailed below.

Nominal Forms. Table 2 shows that most verbs are in the finite or participle form (e.g., "hired"), followed by a smaller proportion in the infinitive. In the gerund, the rate is scarce. However, again, variations in proportions are well controlled.

Verb Tenses. Table 3 highlights that most verbs are in the present tense, followed by much smaller values in the future and the past tenses. Other verb tenses are rare, such as imperfect and pluperfect tenses. Note that the last row highlights cases that do not have verbs (60.8%). For instance, the initial sentence "The mayor of the city of Rio de Janeiro", which is the title of the highlighted act in Figure 1, is a nominal sentence.

Verbal Moods. Table 4 confirms that verbs in the indicative are the ones that appear the most in official diaries since they express facts. Then, there are a few cases of verbs in the subjunctive (i.e., possibility). Conditional verbs (i.e., *if* clauses) and imperative ones (i.e., commands) are pretty rare.

Overall, the syntactic and morphological analyses suggest that the sentences found in administrative acts published in official diaries exhibit consistent and predictable patterns.

3.2 Entity Recognition

This section presents the results obtained with the execution of the entity recognition techniques proposed by Belém *et al*. [2022]. Thus, we aim to identify what indicators arise about recognized entities and their relationships. As the previous

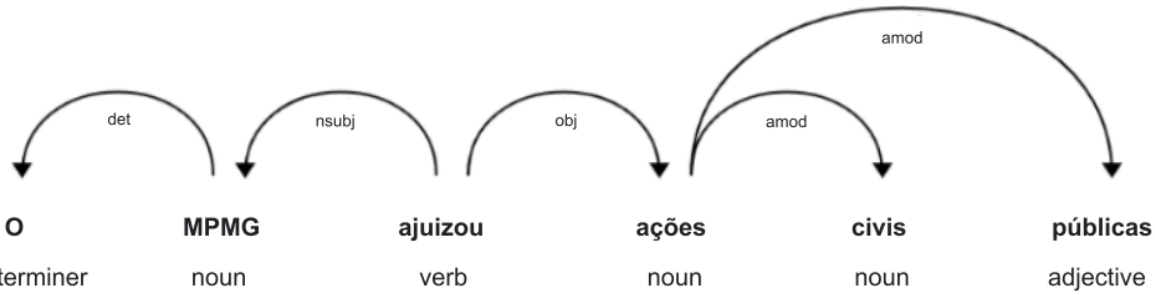

Figure 2. Syntax tree and dependencies between terms generated by the spaCy library. Acronyms: det (article), nsubj (noun subject), obj (object) and amod (noun modifier).

Table 4. Verbal mood present in the sentences.

Mood	Average	Median	SD
Indicative	34.1%	33.8%	4.7%
Subjunctive	1.7%	1.7%	0.6%
Conditional	0.3%	0.3%	0.2%
Imperative	0.0%	0.0%	0.0%
Not Detected	63.9%	64.2%	4.9%

section showed that the official diaries have similar characteristics, we report the results of a random sample of 20 documents.

Named Entity Recognition. This task consists in extracting and classifying entities expressed in natural language (unstructured text) into defined entity types such as *person name*, *location*, *organization*, etc. [Nadeau and Sekine, 2007]. As the official diaries do not have labeled entities or identified relationships, we applied the algorithm proposed by Belém *et al*. [2022] to explain the obtained results, which included the following entity types: *PERSON, ORGANIZATION, TIME, LOCATION, LEGISLATION, JURISPRUDENCE, CPF* (Individual Taxpayer Registration), *CNPJ* (National Registry of Legal Entities)*, TELEPHONE, MASP* (Public Servant Registration Number), and *MONETARY_VALUE*.

Relation Detection. This task involves the characterization of semantic relationships existing between entities. For example, there might be a "works on" relationship between an entity of type *PERSON* and an entity of type *ORGANIZATION*. In addition to relationships between entities, syntactic dependencies between words can also be evaluated. To illustrate this, the sentence *"O MPMG ajuizou ações civis públicas"* (in a literal translation, *"The MPMG judged public civil actions*) generates the syntactic tree shown in Figure 2, indicating how the elements in the clause syntactically relate to each other.

By applying the relation detection algorithm proposed by Belém *et al*. [2022], the relationships between entities tend to reflect commonplace situations regarding legal norms or judicial processes such as *judged by/in*, *published by/in*, *declares*, *granted* and *foreseen*. In particular, Table 5 shows that the most frequent sentence cores in the official diaries are verbs, confirming our previous results shown in Table 1.

In summary, we observed that government acts tend to repeat in official diaries. Thus, in practice, public agencies adopt specific patterns to publicize its acts. For example, in those sections of an official diary related to "Terms of Admission", only the data of the newly hired person is actually changed.

4 Segmentation Process

The segmentation process divides a document into units corresponding to the same action or topic [Pak and Teh, 2018]. The problem is not trivial since most documents deal with different subjects prepared by different instances within the same institution. In this sense, the first step to carry out a segmentation is to preprocess the documents to create a pseudo-structure that will later allow identifying and extracting relevant information.

PDF is the most common format for storing and publishing public acts in official diaries. In general, the snippets to be extracted are well positioned in the presentation layout. Furthermore, the internal structure of a PDF document provides the coordinates for each one of its elements, be they characters or embedded figures, in terms of the X and Y axes. More precisely, due to the unstructured nature of the PDF format, we need to identify not only which characters compose a word, but also which words compose a paragraph, and so on, until we have defined all the elements that compose a single act.

Given this scenario, a strategy to extract information from a given passage is to explore the distances between the textual elements to identify the margins that delimit all the contents belonging to it. The strategy initially converts the PDF document into textual elements with their respective X and Y coordinates. Our initial tests indicated that some documents need to respect the order of appearance of their elements in the presentation. Thus, it is necessary to consider the worst case in which all elements must be sorted according to their coordinates. Once such elements are properly ordered, we define the element types as being *character*, *word*, *line*, and *block*, where lines form a block. A line is made up of words and a word is made up of characters. For each type assigned to an element, there are acceptable horizontal and vertical distance thresholds for deciding which element types will be grouped together or kept apart.

Figure 3 exemplifies this process. In this case, the objective is to determine whether or not the two lines *ESTADO DE MINAS GERAIS* and *PREFEITURA DE ARAÇUAÍ* are part of the same block. To this end, we defined an acceptable threshold for the margin between them (highlighted in red). Similarly, the body and the signature of the snippet are two distinct parts due to the distance defined for the mar-

Table 5. The 14 most significant sentence cores presented in the Brazilian official diaries and their associated frequencies.

Core	Frequency	Core	Frequency
publicada (published)	3,150	período (period)	914
declara (declares)	2,898	lotado (filled)	705
lotada (filled)	2,326	leia-se (read)	691
combinado (combined)	2,165	localizada (localized)	489
publicado (published)	1,410	aposentado (retiree)	392
contar (to count)	1,227	considerando (considering)	352
dada (giving)	1,147	concedidas (granted)	339

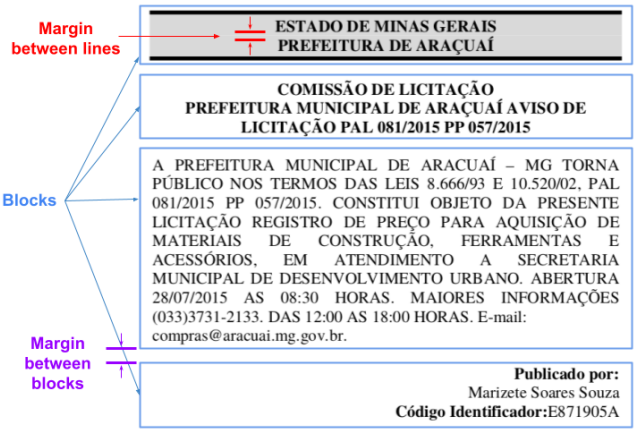

Figure 3. ~~Fragment from an act composed of blocks defined according to the margins between the sub-elements that compose it.~~ **Example of an act structured into blocks defined by the spacing between its sub-elements.**

gin between the blocks (highlighted in purple). Now that the blocks are structured with well-defined parts, each segment in a block corresponds to the information we want to extract.

In addition, it is necessary to address the issue that an excerpt can be displayed on more than one page and consider that there may be a separation of texts in columns on the same page. Such considerations are addressed during the initial sorting process, by adopting a strategy to inspect the number of columns to adjust the relative positions of each chunk as a single stream. In other words, acts that span multiple pages were automatically identified by considering the presence of missing parts (e.g., the signature of an entity such as "Published by") that would be located on the next page. Finally, there is still a step that seeks to discard possible embedded figures (e.g., advertisements or logos) that may appear unexpectedly by repositioning the relative coordinates of the textual elements.

5 Semantic Classification

In this section, we address the problem of receiving as input a set of text segments (public acts) and classifying them according to the semantics found in each one. This classification is essential for more detailed searches for end-point data analysis applications, such as detecting fraud in bids.

The most serious difficulty in this classification task is the large volume of excerpts without any categorization or labeling (700 thousand), which requires manual intervention to annotate them. Moreover, two sub-problems are inherent to the classification process from a repository with only raw texts (i.e., without metadata): (i) dealing with the topic modeling problem, i.e., a statistical model to find which categories/topics the textual excerpts belong to and (ii) defining the proper machine learning technique to learn how to classify the excerpts. We discuss these two sub-problems below and then describe the prototype of a semantic classification tool made available as a Web service.

5.1 Topic Modeling

The identification of semantic topics is a subject related to Natural Language Processing that consists in creating a statistical model to discover abstract topics in a collection of documents [Blei, 2012]. More specifically, topic modeling is a machine learning technique capable of automatically detecting word and phrase patterns in each excerpt of a document to group them according to the similarity of their characteristics.

In our context of public administration, such modeling consists in defining the semantic topics based on the discriminative distribution of the terms that make up each group of excerpts to find topics that are evident, disjoint, and significant. To this end, the *Latent Dirichlet Allocation* (LDA) [Blei *et al.*, 2003] technique[8] was applied, which assumes a Dirichlet probability distribution over textual data to estimate word probabilities for each group. Then, we performed an exploratory analysis by manually varying the number of topics to define the most appropriate number of groups and, thus, to define a semantic topic for each. As a result, four major semantic topics were manually selected: *Personnel*, *Laws*, *Bid*, and *Budget*. We conduct later a more detailed analysis of these specific semantic topics. More specifically, the list of some salient terms (i.e., representative keywords) associated with each semantic topic is as follows:

- Bids: convocação (convocation), edital do pregão (public bidding notice), impugnação do edital (objection to the bidding notice), julgamento de propostas (proposal evaluation), extrato de contrato (contract summary), concorrência (competition), tomada de preço (price quotation), leilão (public auction), Inexigibilidade (unenforceability), and Registro de preço (price registration).
- Laws: decretos (decrees), portarias (ordinances), resoluções (resolutions), circulares (circular letters) and despachos (orders), and regulamentação da lei (regulation of the law).

[8] We note that other specific topic modeling approaches for short texts, such as BTM (*Biterm Topic Models for Short Text*) and *clusterng* (k-means) [Cunha *et al.*, 2021] were also tested. However, the groups were better defined by the LDA.

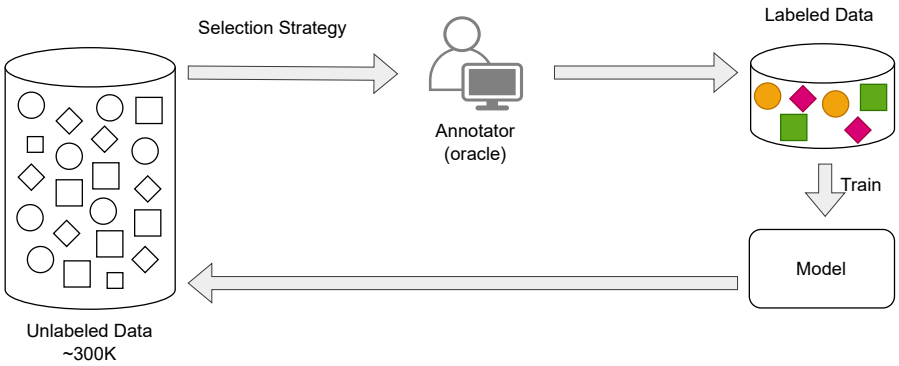

Figure 4. Classification task using active learning.

- Personnel: concurso público (public tender), promoção do emprego (promotion of employment), transferência (transfer of existing job), exoneração (dismissal), demissão (resignation), aposentadoria (retirement), and falecimento (death).
- Budget: planos (plans / arrangements), orçamentos (budgets), diretrizes orçamentárias (budgetary guidelines), prestação de contas (accountability), relatórios de gestão fiscal (fiscal management reports), programação financeira (financial planning), execução orçamentária (budget execution), créditos adicionais (supplementary credits), balanço orçamentário (budget balance), and demonstrativo de receitas e despesas (extract of income and expenses).

5.2 Classification Process

Briefly, the classification problem aims to correctly predict which class a new observation belongs to within the existing possibilities. In the machine learning context, a classification technique is a supervised learning method for correctly attributing new observations to their respective classes, which requires an initial set of already categorized ones (i.e., the training set) [Cunha *et al*., 2021]. After training the classifier, it performs automatically the classification for any new observation. This problem can be addressed by several modeling strategies with different degrees of complexity, such as linear classifiers, *Support Vector Machines* (SVM), decision trees, and neural networks, among others.

Regarding the scenario of public administration, the active learning strategy proves suitable for classifying a large volume of data that initially does not have labels associated with their respective classes [Inuzuka *et al*., 2020; Rangel *et al*., 2020]. This technique consists of repeated interactions between the classifier and an oracle (i.e., a user who classifies specific observations selected by the model). It allows the training of the model to use a reduced number of observations which, when considering a well-done selection process, results in significant learning on the part of the classifier with a low cost of labeling [Lewis and Catlett, 1994].

The selection of observations that the oracle will consult can be made differently. For example, one option would be to consult the oracle about observations close to class boundaries, as the model would find it more difficult to classify them. However, combining these cases with instances in which the model has some confidence to classify may be interesting, since this consolidates the classes already determined by the classifier.

Thus, choosing a reduced number of more significant samples for training the classifier (the most helpful sample) depends on the adopted strategies. Here, such strategies were based on *classification uncertainty* [Lewis and Catlett, 1994]. Therefore, they are called uncertainty measures. The adopted approach considers three integrated measures: *classification uncertainty, classification margin* and *classification entropy*. Classification uncertainty is the most straightforward one, since it selects the sample based on this rank measure, which means selecting the one with the highest uncertainty. Yet, the classification margin selects the sample with the smallest associated distance from the decision boundary. Finally, regarding the classification entropy measure, it is proportional to the average number of guesses one must make to find the valid class. The more uniform the distribution, the greater the entropy; therefore, the most uncertain sample is the one with a high entropy value. As detailed later in Section 6.2, we adopted the uncertainty strategy as the selection algorithm, which aligns with studies on text classification of official documents [Inuzuka *et al*., 2020; Rangel *et al*., 2020].

Figure 4 illustrates the process of labeling and classifying segments in official diaries. First, based on a selection strategy (uncertainty, margin, or entropy samples), the tool selects a set of unlabeled data from the repository that contains approximately 300,000 segments. Next, the oracle identifies and labels these selected segments forming the training set. Finally, the supervised learning module receives the training set, in which each example, represented by its characteristics or attributes, is labeled according to the class to which it belongs. As a result, the learning system must build and update the model, which allows it to predict classes for new inputs different from previously labeled examples, restarting the machine learning cycle.

Initial results showed that classes are naturally unbalanced. Hence, we addressed this issue by adopting three strategies: (i) prioritizing minority classes in the strategy for selecting segments to be labeled; (ii) dealing with imbalance by reducing the number of observations of the majority class (i.e., undersampling) directly in the model training phase; and (iii) promoting pseudo-labeling to balance the class using a semi-supervised strategy as described later in Section 6.1.

5.3 Active Learning Process as a Web Service

Although there are good open-source tools for data annotation, like Brat[9] and Doccano[10], they do not support machine learning techniques that require real-time interactivity with the annotator. The best alternatives are commercial and closed source, like the Prodigy tool[11]. To overcome this problem, we propose an annotation tool integrated with active learning as a Web service. Technically, the proposed solution follows a pattern similar to closed-source alternatives, which has the main advantage of easy integration with predictive models, whether for local applications, network applications, or even mobile devices.

Figure 5 illustrates our proposal, an Annotation tool with an active learning support. Its architecture consists of three modules: a Web Interface (*front-end*), a Controller (*middleware*), and an Active Learning (*back-end*). This setup enables users to create interactive notes, while the *Controller* carries out a predefined set of GET/POST requests to the Active Learning module. Furthermore, since each module is independent, the annotator tool can transparently couple with new implementations of predictive models and selection strategies.

Besides, Figure 5 also exemplifies the user interface for the real-time interactive annotation process built into the active learning model. The graph shows the model's accuracy as the annotator decides which label is most appropriate for a given sample. Furthermore, this implementation allows multiple users to annotate the same set of instances simultaneously. Therefore, the predictive processing is all performed in real-time and can be finished as soon as the accuracy is satisfied.

6 Experimental Setup

In this section, we first describe the segmentation process that generates the datasets used in our experiments (Subsection 6.1). Subsequently, we briefly discuss the classification strategies that have been employed (Subsection 6.2). Finally, we provide details about the metrics and measures used in our classification task (Subsection 6.3).

6.1 Data: Segmentation Process

As input for the segmentation process, 1,640 documents in PDF format were considered. As output, the segmentation process extracted 645 Federated Entities and 307,277 segments. It should be noticed that an official diary addresses distinct Federated Entities and, in turn, each one of them can have several sections (i.e., acts under its responsibility). Table 6 shows basic statistics about the federated entities extracted per document.

Given that part of the segmentation process consists in ordering the textual elements according to their coordinates on each page (with complexity equal to $O(n \log n)$), the analysis of the total processing time of real cases is of major importance. The graph in Figure 6 shows the distribution of document sizes, where the Y axis refers to the proportion of files in PDF format and the X axis refers to file sizes. It is observed that 93% of all documents are less than 5 MB in size. The processing of the largest document (20.1 MB with 667 pages) took 12 minutes on an Intel Duo CPU 2.6 GHz machine with 4 GB of RAM memory, which is quite acceptable given the daily frequency of documents published by each federal entity.

Table 6. Statistics of the data present in the 1640 documents.

	# of Entities per document	# of Segments per document
Average	96.9	426.7
Median	92	416.5
Standard Deviation	39.5	190

Table 7. Proportion of the semantic classes of the Official Diaries excerpts.

Class	Proportion
Bid	58.96%
Personnel	24.77%
Laws	10.03%
Budget	6.23%

The quality of the extractions was manually assessed considering 5,219 randomly selected samples using the annotation prototype (i.e., it is informed whether or not the segment was properly collected), being verified that the segments were correctly extracted (i.e., the accuracy was of 100%). On the other hand, there are minor imperfections, such as elements coming from tables slightly out of position. However, the segments' textual content was fully preserved to be used in future stages of classification and textual search. Therefore, the structure-based strategy generates a well-defined pattern according to the distances between the textual elements. In particular, the strategy proved to be an excellent technical choice to deal with possible out-of-order formatting, which it was a recurring issue during our testing.

As already mentioned, approximately 300,000 unlabeled excerpts (i.e., government acts) must be classified appropriately. For this, the active learning process requires an initial set to make its first decisions. In this sense, 29 initial examples were randomly selected, separating an initial set with ten instances and the test set with the remaining 19. Then, we performed the annotation process according to the strategy that selects the sample with greater uncertainty about it, that is, the most helpful example to be labeled. This way, 658 textual excerpts were manually labeled (equivalent to 0.2% of the repository). As a result, we observed that the nature of the problem is highly unbalanced, with the subject related to *Bid* being much more representative (62.45%) than the ones referring to *Laws* (10.03%) and *Budget* 6.23%, as shown in Table 7.

6.2 Models: Semantic Classification

We now present the experimental evaluation of the semantic classification process carried out using active learning, following the strategy proposed in Section 5. Specifically, we selected four state-of-the-art classifiers as predictive models: (i) **SVM**, widely used for text classification, having produced the best results in a recent benchmark [Cunha

[9] Brat, accessed 22 June 2023, `https://brat.nlplab.org`.

[10] Doccano, accessed 22 June 2023, `https://doccano.github.io/doccano`.

[11] Prodigy, accessed 22 June 2023, `https://prodi.gy`.

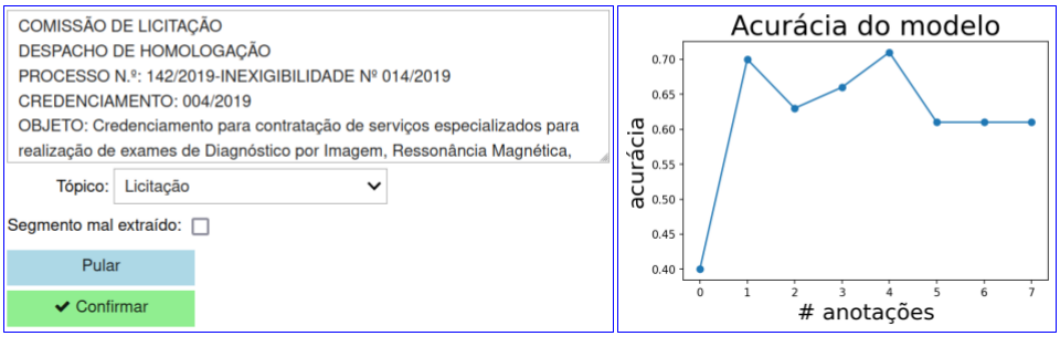

Figure 5. Real-time interaction interface with the annotator integrated into the active learning module.

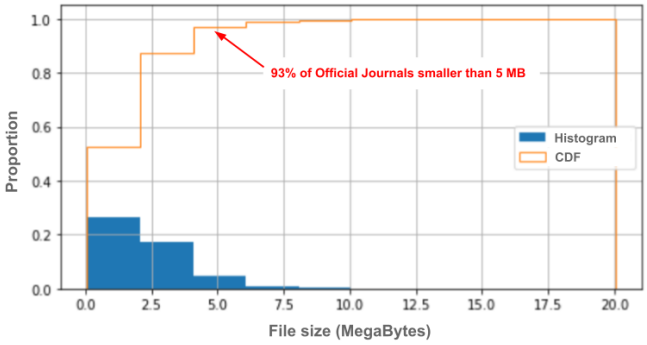

Figure 6. Distribution of document sizes.

et al., 2021] in scenarios similar to those described here, that is, small or medium collections, with little labeled data and high-class imbalance; (ii) **BERT** [Devlin *et al*., 2019], which has been the state-of-the-art in numerous NLP tasks [Cunha *et al*., 2021; Garg *et al*., 2020]; (iii) **BERTimbau** [Souza *et al*., 2020], which extends BERT to recognize sentences in Brazilian Portuguese; and (iv) **LaBSE** [Feng *et al*., 2022], which explores the use of language models capable of representing sentences in several languages. Regarding the algorithm selection techniques in the active learning phase, the uncertainty strategy [Lewis and Catlett, 1994] obtained the best results. Such algorithms are in line with other initiatives for classifying texts extracted from official diaries [Inuzuka *et al*., 2020; Rangel *et al*., 2020].

To increase the robustness of the classification strategies mentioned in this section, we also added two new strategies: *stacking* and *combination of representations*. Stacking exploits the prediction information from different base classifiers by means of a meta-classifier fed with the predictions of the base ones [Džeroski and Ženko, 2004; Wolpert, 1992]. This strategy combines the effectiveness of different classifiers, exploiting their complementarities with the goal of surpassing the individual classifier's results. Recent works that exploit this strategy along with Transformers, as well as more traditional classifiers such as SVM, demonstrate the effectiveness of this approach in text classification tasks similar to the ones we face here [Campos *et al*., 2017a; Gomes *et al*., 2021; Salles *et al*., 2018]. Likewise, we use SVM, BERTimbau and LaBSE as the base models for stacking and feeding a meta-classifier – Logistic Regression – with information (prediction and confidence) outputted by the three aforementioned models.

In the combination of representations strategy, a textual document is represented by different forms (for example, TF-IDF, Word Embedding, Text graph). These different representations are used used to train a classification model. Recent works in the literature [Carvalho and Plastino, 2021; Córdova Sáenz *et al*., 2021; de Andrade, Claudio Moisés Valiense and Gonçalves, Marcos André, 2020] combined such representations with the aim of improving the effectiveness of the classification task. This strategy differs from the stacking one as it does not learn how to combine the output of different classifiers, but instead, directly exploits the complementary information existing in each representation. In here, we perform a combination by simply concatenating the vector representations of the documents based on TF-IDF (frequentist), FastText (based on static embeddings), and BERTimbau (based on contextual embeddings).

6.3 Evaluation

To evaluate the effectiveness of the semantic classification task, we first consider *accuracy* that measures the global effectiveness of all decisions made by the classifier (hit ratio). We also use *Precision* that informs how many classes are correct among those associated with a given class and *recall* that indicates which classes were correctly associated concerning the total number of instances of a given class. Finally, the *F1-Measure* corresponds to the harmonic mean between precision and recall [Cunha *et al*., 2021]. In the case of the F1-Measure, we present the results of the Macro (F1) variant that indicate its averages per class, thus being more suitable for highly unbalanced problems, which is the case here.

To evaluate the robustness of the semantic classification procedures, we apply a *folded cross-validation procedure* in our experiments. This experimental technique explores the training of several models in different subsets of input data (training sets) and the evaluation of them in the complementary subsets of the data (test sets) [Cunha *et al*., 2021]. One of the primary purposes of a cross-validation is to verify the generalization of the models for different training and test sets. To choose the best parameters of the classifiers, we used cross-validation within the training (*nested cross-validation*).

Finally, to deal with issues of (labeled) data scarceness and dataset skewness, we exploit an oversampling strategy, in which we applied a data augmentation approach to the training sets to smooth the large class imbalance. Specifically, we first build a model using BERTimbau, one of the most effective classification algorithms, as we shall see, in each of the five runs of the five fold cross-validation. Next, we used this model to predict the class of other unla-

Table 8. Accuracy, Macro-F1 and confidence interval of 95%.

Method	Diarios		Diarios+	
	Accuracy	**Macro-F1**	**Accuracy**	**Macro-F1**
SVM	0.86 ± 0.02	0.73 ± 0.07	0.87 ± 0.02	0.75 ± 0.07
FastText	0.81 ± 0.03	0.63 ± 0.03	0.72 ± 0.02	0.37 ± 0.02
BERT	0.84 ± 0.02	0.71 ± 0.05	0.84 ± 0.01	0.72 ± 0.04
BERTimbau	0.86 ± 0.01	0.75 ± 0.05	0.86 ± 0.04	0.76 ± 0.07
LaBSE	0.86 ± 0.03	0.77 ± 0.06	0.86 ± 0.03	0.77 ± 0.06
Concat (TF-IDF, FastText)	0.85 ± 0.03	0.71 ± 0.07	0.86 ± 0.01	0.75 ± 0.03
Concat (TF-IDF, BERTimbau)	0.84 ± 0.02	0.74 ± 0.06	0.84 ± 0.03	0.74 ± 0.06
Stacking	0.88 ± 0.03	0.79 ± 0.09	0.85 ± 0.02	0.74 ± 0.05

Table 9. Precision (P), Recall (R) and F1-score (F1) of classifiers by class. Confidence intervals are omitted for space reasons.

Models		Diários				Diários+			
		Law	Bid	Budget	Personnel	Law	Bid	Budget	Personnel
SVM	P	0.62	0.87	0.95	0.85	0.63	0.89	0.95	0.84
	R	0.15	0.96	0.88	0.9	0.22	0.96	0.88	0.92
	F1	0.22	0.91	0.91	0.87	0.31	0.92	0.91	0.88
BERT	P	0.36	0.88	0.92	0.81	0.43	0.88	0.98	0.82
	R	0.18	0.96	0.83	0.82	0.21	0.96	0.83	0.82
	F1	0.24	0.91	0.87	0.82	0.26	0.92	0.89	0.82
BERTimbau	P	0.52	0.88	0.92	0.88	0.52	0.89	0.93	0.87
	R	0.23	0.95	0.83	0.92	0.29	0.95	0.81	0.9
	F1	0.32	0.91	0.87	0.9	0.35	0.92	0.86	0.88
LaBSE	P	0.5	0.88	0.92	0.87	0.66	0.89	0.89	0.89
	R	0.27	0.94	0.88	0.9	0.3	0.95	0.88	0.87
	F1	0.35	0.91	0.9	0.88	0.38	0.92	0.88	0.88
C(TF-IDF, BERTimbau)	P	0.44	0.89	0.85	0.87	0.45	0.89	0.85	0.86
	R	0.28	0.93	0.86	0.84	0.3	0.93	0.86	0.84
	F1	0.33	0.91	0.85	0.85	0.34	0.91	0.85	0.85
Stacking	P	0.58	0.89	0.89	0.88	0.5	0.89	0.95	0.86
	R	0.31	0.95	0.88	0.9	0.27	0.94	0.88	0.88
	F1	0.4	0.92	0.87	0.89	0.31	0.92	0.91	0.87

beled segments, resulting in about 300.000 pseudo-labeled segments. Finally, we randomly selected pseudo-labeled segments to add to the smaller classes until all classes had an equal number of samples equivalent to the size of the Bid (majority) class. We call this version of the dataset *Diarios+*.

We should stress that these procedures are applied only on the training split of each fold while the test splits were kept untouched in all cases, meaning that the results of both strategies (with and without oversampling) are directly comparable.

7 Experimental Results

We report a comprehensive experimental evaluation of the classification task by considering traditional models, transformers, stacking, and the combination of representations. Table 8 shows the average results of Accuracy (*aka* MicroF1) and MacroF1 on the test *folds* of the cross-validation process with 5 folds (5-fold cross-validation procedure) in both datasets, without and with oversampling: *Diarios* and *Diarios+*. Particularly, in the FasText line, we use word embedding for the Portuguese language[12]. In the line Concat (TF-IDF, FastText), given a document X represented by the vector Y in the TF-IDF representation and Z in the FastText representation, we concatenate these vectors, resulting in a single representing for a document. We perform a similar process for Concat (TF-IDF, BERTimbau).

As we can see, the accuracy of all classifiers is similar (in statistical terms), being all quite effective (around 85%). On the other hand, the results of MacroF1, which is a more adequate measure for problems with class imbalance, show that the Bertimbau and LaBSE figures are considerably higher than those of SVM (2.6 - 5.2% higher) and BERT (5.3 - 7.8% higher). Notably, the MacroF1 of both (LabSE and BERTimbau) is relatively high, around 75%.

When comparing the results of both strategies to deal with the skewness, we see no real advantage of one over the other, as the results of all methods, considering both metrics, are statistically tied, with the exception of the MacroF1 of FastText, which is much worse in *Diarios+*. Remind that despite the fact that the oversampling strategy potentially increases the data sample, which could benefit mainly the Transformers, this strategy may also introduce noise in the pseudo-labels used by classifiers to build the models, decreasing their effectiveness. Though this hypothesis needs further investigation, it also could explain the poor effectiveness of Fasttext in *Diarios+* as the embeddings created by this method are more sensitive to the introduction of noise in the labeling.

Considering a per class evaluation, Table 9 shows Preci-

[12] FastText, Word vectors for 157 languages, 22 June 2023, `https://fasttext.cc/docs/en/crawl-vectors.html`.

sion, Recall, and F1-Measure per class for the best strategies shown in the Table 8, for both datasets[13]. As expected, the majority class (Bid) has the highest values for all metrics (between 87% and 94%) for all classifiers. Despite a statistical tie with all other methods in Table 8, Stacking is the classifier that achieves the most "balanced" performance for the four classes, mainly for the minority ones (e.g., Budget) in the *Diarios* dataset. One explanation for this case is the capacity of stacking to take advantage of the opinion of several classifiers and correct mistakes. Regarding the base classifiers, LaBSe and Bertimbau stand out also in this analysis. Finally, regarding the comparison of the strategies without and with oversampling, the latter seems to benefit a bit some strategies in terms of recall, but not enough to produce a statistically significant improvements according to the Wilcoxon test.

In sum, despite testing several advanced techniques (stacking, combination of representations) in the literature, we have been unable to surpass (statistically speaking) the results reported in our previous work [Constantino *et al*., 2022], not even with the new oversampling strategy, which probably means that our current solutions have achieved the limits of what can be achieved with the current state-of-the-art in automated text classification techniques. But if we have to suggest a strategy, we would recommend to use Stacking without oversampling, if cost is not an issue, or LabSe alone, also without oversampling, if cost is a factor.

Regarding classification errors, we observed two main types: *annotation error* and *prediction error* of the model. For the first case, we can illustrate it with the excerpt[14] *"Term of Amendment to the Contract [anonymous], Process [anonymous], for the acquisition of..."*, with the title *"ORDINANCE/LAWS"*, which was manually labeled as *Law*, but is a specific case of *Bid*. In this case, there is information in the title of the excerpt that is related to the categories to which it would fit (i.e., ordinance or law), leading the annotator to make a mistake. However, carefully analyzing the content, we notice that it is an amendment to a bidding contract.

For the second case, the SVM model, for instance, erroneously predicted as belonging to the *Personnel* class the fragment of the excerpt *"ORDINANCE [anonymous] REGULATES THE COMPLEMENTARY LAW [anonymous] THAT INSTITUTES THE STAFF OF ..."*[15], which corresponds to a *law regulation*. This error example illustrates the difficulty of classifying excerpts with ambiguity, even for a human evaluation. In this case, the algorithm considered the term "personnel staff" as strongly associated with the *Personnel* class.

8 Conclusions and Future Work

This article addresses a real problem in terms of preprocessing and organization of public documents as part of a collaboration established between the Department of Computer Science of the Federal University of Minas Gerais and the Public Ministry of the State of Minas Gerais. More specifically, the article addresses critical aspects involving the segmentation and the semantic classification of official diaries excerpts.

As a solution, we propose to semantically classify extracted excerpts from official journals with an active learning strategy that minimizes manual labeling effort. Furthermore, we implemented a data labeling tool as a web service integrated into the active learning process. Regarding semantic classification, with very little labeling effort and adopting techniques to deal with the natural imbalance of semantic classes, our cross-validation evaluations indicated a value for *Macro-F1* of 75% and an overall accuracy of 85%. As part of our extended work, we employed morphosyntactic analysis and entity recognition to characterize excerpts from official diaries, which may be useful as meta-features in future research. To deal with issues of (labeled) data scarceness and skewness, we also exploited new oversampling strategies based on semi-supervised learning to generate pseudo-labeled samples.

Concerning new contributions to the classification task, we sought to enhance the semantic classification process by incorporating two additional strategies: *stacking* and *combination of representations*. Specifically, we used the SVM, BERTimbau and LaBSE models to implement a stacking technique, while we concatenated document´s vector representations based on of TF-IDF (frequentist), FastText (static embeddings) and BERTimbau (contextual embeddings) to combine the representations´strengths. To deal with issues of (labeled) data scarceness and skewness we also propose to explore an oversampling strategy using semi-supervised techniques. Our experimental results indicate that, despite all efforts, we were unable to produce statistically significant improvements over our previous results, which probably means that we are close to the limits of what can be achieved in terms of classification effectiveness with the current state-of-the-art in automatic text classification. This also means that our current solutions remain state-of-the-art in the field of semantic classification considering the Brazilian official diary context. In practice, associating labels to acts enriches the indexing of documents to help organize information such as fraud detection, information retrieval and transparency. For instance, Belém *et al*. [2022] employed our extraction strategies to enhance their entity recognition algorithm.

As part of our future work, we aim to conduct an investigation into the annotation process, specifically focusing on use cases that involve active learning with stakeholders, i.e., exploring qualitative aspects with the goal of assessing the quality of the trained dataset and monitoring the time required to accomplish specific tasks. Additionally, we intend to expand the source of government documents beyond the scope defined by our current partnership with the Public Ministry of the State of Minas Gerais (MPMG) and address, particularly, difficult document cases that have embedded images (e.g., scanned documents). We also intend to apply active learning to end-point applications, for example, to decide whether or not there is evidence of fraud in acts published in the official journals. Finally, we intend to further study the cases of fail-

[13] We omit confidence intervals for space reasons.

[14] A literal translation of the content *"Termo de aditamento ao Contrato [anonymous], Processo [anonymous], para aquisição de ..."* with the title "PORTARIAS/LEIS".

[15] A literal translation of *"PORTARIA [anonymous] REGULAMENTA A LEI COMPLEMENTAR [anonymous] QUE INSTITUI O QUADRO PESSOAL DA ....*

ure in classification with the goal of boosting even further the effectiveness of our **segmentation and classification strategies**, especially for underrepresented classes (e.g., Law).

Acknowledgements

We would like to thank the Public Ministry of the State of Minas Gerais for its support under the *Capacidades Analíticas* project.

Funding

Work partially supported by the authors' individual grants from CAPES, CNPq, FAPEMIG and Fundação Araucária.

Authors' Contributions

Kattiana Constantino and Thiago H. P. Silva: Conceptualization, Writing (review & editing), Methodology, Validation. *Marcos Carvalho, Victor Augusto L. Cruz, and Otávio M. M. Zucheratto*: Writing (review & editing), Methodology, Validation. *João Vítor B. Silva, Welton Santos, Celso França, and Claudio M. V. de Andrade*: Writing (review & editing), Validation. *Alberto H. F. Laender and Marcos A. Gonçalves*: Conceptualization, Writing (review & editing), Validation, Project Management, Supervision.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets and the source code of the current work are available at `https://github.com/MPMG-DCC-UFMG/M02`.

References

Belém, F. M., Ganem, M., França, C., Carvalho, M., Laender, A. H. F., and Gonçalves, M. A. (2022). Reforço e Delimitação Contextual para Reconhecimento de Entidades e Relações em Documentos Oficiais. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 292–303, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2022.224650.

Blei, D. M. (2012). Probabilistic Topic Models. *Commun. ACM*, 55(4):77–84. DOI: 10.1145/2133806.2133826.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Campos, R., Canuto, S., Salles, T., de Sá, C. C., and Gonçalves, M. A. (2017a). Stacking Bagged and Boosted Forests for Effective Automated Classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 105–114, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3077136.3080815.

Campos, R. R., Canuto, S. D., Salles, T., de Sá, C. C. A., and Gonçalves, M. A. (2017b). Ranked Batch-mode Active Learning. *Inf. Sci.*, 379:313–337. DOI: 10.1016/j.ins.2016.10.037.

Carvalho, J. and Plastino, A. (2021). On the Evaluation and Combination of State-of-the-Art Features in Twitter Sentiment Analysis. *Artif. Intell. Rev.*, 54(3):1887–1936. DOI: 10.1007/s10462-020-09895-6.

Constantino, K., Cruz, V. A. L., Zucheratto, O. M. M., França, C., Carvalho, M., Silva, T. H. P., Laender, A. H. F., and Gonçalves, M. A. (2022). Segmentação e Classificação Semântica de Trechos de Diários Oficiais Usando Aprendizado Ativo. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 304–316, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2022.224656.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S. D., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J. M., Rosa, T., Rocha, L., and Gonçalves, M. A. (2021). On the Cost-Effectiveness of Neural and Non-Neural Approaches and Representations for Text Classification: A Comprehensive Comparative Study. *Inf. Process. Manag.*, 58(3):102481. DOI: 10.1016/j.ipm.2020.102481.

Córdova Sáenz, C. A., Dias, M., and Becker, K. (2021). Assessing the Combination of DistilBERT News Representations and Difusion Topological Features to Classify Fake News. *Journal of Information and Data Management*, 12(1). DOI: 10.5753/jidm.2021.1895.

de Andrade, Claudio Moisés Valiense and Gonçalves, Marcos André (2020). Combining Representations for Effective Citation Classification. In *Proceedings of the 8th International Workshop on Mining Scientific Publications*, pages 54–58, Wuhan, China. Association for Computational Linguistics.

de Freitas, J., Pappa, G. L., da Silva, A. S., Gonçalves, M. A., de Moura, E. S., Veloso, A., Laender, A. H. F., and de Carvalho, M. G. (2010). Active Learning Genetic Programming for Record Deduplication. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2010, Barcelona, Spain, 18-23 July 2010*, pages 1–8. IEEE. DOI: 10.1109/CEC.2010.5586104.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics. DOI: 10.18653/v1/n19-1423.

Džeroski, S. and Ženko, B. (2004). Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, 54:255–273. DOI: 10.1023/B:MACH.0000015881.36452.6e.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 878–891. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.62.

Fernandes, D., de Moura, E. S., Ribeiro-Neto, B., da Silva, A. S., and Gonçalves, M. A. (2007). Computing Block Importance for Searching on Web Sites. In *Proceedings of the sixteenth ACM conference on Conference on infor-*

mation and knowledge management, pages 165–174. DOI: 10.1145/1321440.1321466.

Garg, S., Vu, T., and Moschitti, A. (2020). TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 7780–7788. AAAI Press.

Gomes, C., Gonçalves, M. A., Rocha, L., and Canuto, S. D. (2021). On the Cost-Effectiveness of Stacking of Neural and Non-Neural Methods for Text Classification: Scenarios and Performance Prediction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4003–4014, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2021.findings-acl.350.

Inuzuka, M., do Nascimento, H., Almeida, F., Barros, B., and Jradi, W. (2020). Doclass: Open-source Software to Support Document Labeling and Classification. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*, pages 105–112, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/kdmile.2020.11965.

Knackstedt, R., Heddier, M., and Becker, J. (2014). Conceptual Modeling in Law: An Interdisciplinary Research Agenda. *Communications of the Association for Information Systems*, 34(1):36. DOI: 10.17705/1cais.03436.

Lewis, D. D. and Catlett, J. (1994). Heterogeneous Uncertainty Sampling for Supervised Learning. In *Machine Learning Proceedings 1994*, pages 148–156. Elsevier. DOI: 10.1016/b978-1-55860-335-6.50026-x.

Lin, F.-R., Chou, S.-Y., Liao, D., and Hao, D. (2015). Automatic Content Analysis of Legislative Documents by Text Mining Techniques. In *2015 48th Hawaii International Conference on System Sciences*, pages 2199–2208. IEEE. DOI: 10.1109/HICSS.2015.263.

Nadeau, D. and Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, 30(1):3–26.

Pak, I. and Teh, P. L. (2018). Text Segmentation Techniques: A Critical Review. *Innovative Computing, Optimization and Its Applications*, pages 167–181. DOI: 10.1007/978-3-319-66984-$7_1$0.

Pereira, G. C., Monteiro, I. T., Vasconcelos, D. R., Braz, L., and Silva, C. H. (2021). Classificação Taxonômica de Categorias de Serviços Públicos para Aplicações Digitais. In *Anais do IX Workshop de Computação Aplicada em Governo Eletrônico*, pages 119–130, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wcge.2021.15982.

Pinto, F., Santos, J., Lifschitz, S., and Haeusler, E. (2023). A Benchmarking for Public Information by Machine Learning and Regular Language. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 60–71, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wcge.2023.229975.

Pinto, F. A. D., Haeusler, E. H., and Lifschitz, S. (2021). Transparência Pública Automatizada a Partir da Gramática do Diário Oficial. In *Anais do IX Workshop de Computação Aplicada em Governo Eletrônico*, pages 59–70, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wcge.2021.15977.

Rangel, M., Bernardini, F., Viterbo, J., Monteiro, R., Seixas, E., and dos Santos Pinto, H. (2020). Uso de Aprendizado de Máquina para Categorização Automática de Conjuntos de Dados de Portais de Dados Abertos. In *Anais do VIII Workshop de Computação Aplicada em Governo Eletrônico*, pages 120–131, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wcge.2020.11263.

Salles, T., Gonçalves, M. A., Rodrigues, V., and Rocha, L. (2018). Improving Random Forests by Neighborhood Projection for Effective Text Classification. *Information Systtems*, 77:1–21. DOI: 10.1016/j.is.2018.05.006.

Silva, M. O., Costa, L. L., Bezerra, G., Gomide, L. D., Hott, H. R., Oliveira, G. P., Brandão, M. A., Lacerda, A., and Pappa, G. (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.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems, (BRACIS)*, pages 403–417. Springer.

Wolpert, D. H. (1992). Stacked Generalization. *Neural Networks*, 5(2):241–259. DOI: https://doi.org/10.1016/S0893 - 6080(05)80023-1.