HALF: Human-Assisted Labeling Feedback Method for Subject Mining

Bruno Rogério S. dos Santos[1], Leonardo Sampaio Rocha[1], Vinícius de M. Menezes[1], Evilásio Costa Júnior[2], Pedro Henrique L. Lessa[1]

[1]LAGIC, State University of Ceará
Fortaleza – CE – Brazil

[2]Department of Computer Science, Federal University of Ceará
Sobral – CE – Brazil

bruno.rogerio@aluno.uece.br

leonardo.sampaio@uece.br, vinicius.menezes@aluno.uece.br

junior.facanha@gmail.com, pedro.luna@aluno.uece.br

Abstract. *The increasing volume of unstructured texts in Official Gazettes highlights the need for robust semantic search engines. To address this, we propose a hybrid approach combining machine learning, human supervision, and a Large Language Model (LLM). The HALF (Human-Assisted Labeling Feedback) method, leveraging GPT-4o Mini, classified subjects in publications from the Official Gazette of Ceará. It assigned subjects to 1.044 publications with 0.8889 accuracy compared to ground truth. This approach enhances semantic search, improves retrieval and decision-making, and extends to other legal domains. Moreover, it offers a scalable solution, outperforming traditional unsupervised methods in accuracy and relevance.*

1. Introduction

The increasing volume of publications frequently published in Official Gazettes, containing large and unstructured texts, highlights the need for robust semantic search engines. However, challenges arise due to the diverse formats and structures of these documents, requiring efficient processing methods. Ensuring accessibility to this data benefits the public, researchers, and practitioners by maintaining transparency and monitoring public resources. For example, NLP techniques can track the use of public funds, detecting suspicious spending patterns and contracts [Pinto et al. 2023].

To address these challenges, this paper presents Human-Assisted Labeling Feedback (HALF), a method that combines human supervision with a large language model (LLM) to accurately classify publication subjects. By applying data preprocessing techniques such as normalization, tokenization, and vectorization, the method improves the definition of relevant subjects for official publications. Our method facilitates searches in the Official Gazettes of Ceará by classifying publications based on their content.

We applied our method to classify 1.044 publications over two years, comparing the results with labels from the Secretariat of Management Planning (SEPLAG). Our method achieved an accuracy of 0.8889. A key advantage of HALF is its scalability, allowing it to be applied to journals and legal documents from other states, including the Official Gazette of the Union.

2. Background

2.1. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of artificial intelligence focused on enabling machines to understand, interpret, and generate human language. Text preprocessing plays a crucial role in NLP and machine learning, ensuring data quality and consistency. This includes cleaning steps such as noise removal, normalization, and dimensionality reduction techniques like stop-word removal and tokenization, which break text into meaningful units for analysis [Eisenstein 2018].

Effective preprocessing enhances model accuracy by improving data consistency and computational efficiency. Techniques such as lemmatization help neural networks process text more effectively, reducing complexity and training time. Additionally, methods like TF-IDF and Word Embeddings refine keyword extraction, while Bag-of-Words and deep learning models (e.g., Word2Vec, BERT) create richer text representations.

Ultimately, text preprocessing is essential for transforming raw textual data into structured and interpretable formats. By improving data quality, efficiency, and model accuracy, it lays the foundation for reliable NLP applications. Investing in preprocessing techniques ensures significant gains and optimal performance in text-based projects [Christopher et al. 2008].

2.2. Text Similarity with Hierarchical Analysis

Text similarity is a fundamental problem in NLP with applications in information retrieval, text clustering, and document summarization. Traditional methods such as TF-IDF, cosine similarity, and word embeddings provide lexical and semantic comparisons but struggle to capture hierarchical relationships. This limitation is particularly relevant in tasks like document categorization and knowledge organization, where understanding structural dependencies between text components is crucial [Jurafsky and Martin 2025].

Hierarchical similarity analysis enhances text comparison by incorporating structured relationships between textual elements. This approach recognizes that sentences and paragraphs contribute to broader thematic constructs, allowing for a more nuanced assessment of content relevance. By applying hierarchical strategies, it becomes possible to refine similarity scores, distinguishing core content from peripheral details, especially in domains such as legal and governmental documents.

Recent advances in deep learning algorithms, such as BERT and Sentence Transformers, improve text similarity through contextual embeddings. Integrating these models with hierarchical techniques refines similarity assessments, ensuring better alignment between topic classifications. This study applies hierarchical similarity to validate whether subjects in the HALF method encompass or align with those in SEPLAG, enhancing the accuracy and reliability of NLP-driven topic classification [Zangari et al. 2024].

2.3. Large Language Model (LLM)

Large Language Models (LLMs) are advanced neural networks designed for large-scale natural language processing. Models such as GPT-4 and LLaMA3 are trained on vast textual datasets, enabling them to perform tasks like text summarization, translation, and

question answering. With billions of parameters, these models capture complex linguistic structures and contextual nuances.

LLMs utilize deep learning, particularly the Transformer architecture, which revolutionized NLP by enabling parallel text processing. Key mechanisms include tokenization, attention mechanisms for contextual understanding, and positional encoding to preserve word order. Training involves a pre-training phase, where models learn general language patterns, followed by fine-tuning on specific tasks like sentiment analysis and legal text classification [Vaswani 2017].

These models represent a breakthrough in NLP, expanding automation across sectors, including legal and governmental applications. With ongoing improvements, LLMs continue to enhance text understanding, ensuring more accurate and context-aware language processing [Brown et al. 2020].

3. Related Works

In this section we present some works related to ours, which use machine learning techniques to process and classify official documents.

In [Guimarães et al. 2024] an automated tool developed for Named Entity Recognition (NER) in Official Gazettes of the Federal District (OGFD) is presented, with the aim of facilitating the search and retrieval of information in government documents. Its main strengths include the combination of rule-based methods and machine learning, which improves accuracy in data extraction. The tool implements text segmentation through regular expressions, optimizing the organization of acts present in diaries. Focused on Brazilian legal documents, DODFMiner offers a solution accessible through a command line interface (CLI) and can be easily installed via Python Package Index (PyPi). Furthermore, it has detailed documentation that guides users in its use. Trained with high-quality corpora, the tool provides simple and accurate results in extracting named entities, contributing to public transparency.

This article [Castano et al. 2024] discusses the importance of extracting legal information and presents the ASKE approach to this, which uses advanced NLP and machine learning techniques. The objective is to improve the analysis and understanding of legal documents, which is the case of this article, more specifically legal documents. ASKE (Advanced Semantic Knowledge Extraction) is a technique that aims to extract relevant information from legal documents, considering the context in which the words appear. In addition to all this, the article mentions the use of contextual embeddings, which take into account the full text of a word to improve meaning disambiguation. The approach also combines external resources, such as ontologies, with NLP techniques for more robust semantic analysis. Zero-shot learning (ZSL) is highlighted as a promising solution, enabling never-before-seen class-specific classifications during training.

The work [Dobša and Kiers 2022] presents an approach to document classification through semi-supervised clustering in a semantic space. Its strength lies in leveraging both labeled and unlabeled data to create clusters that improve the accuracy of document classification. By mapping documents into a semantic space, the model can group similar texts together, improving the classification of documents with little labeled data. The method also benefits from the use of semantic enrichment, which enhances the clustering process by making better use of context.

The approach presented in [Cação et al. 2021] consists of applying a deep learning model, specifically BERT, to classify and monitor government acts published in the Official Gazette of the Union of Brazil, with a focus on environmental policies. One of the main strengths of this approach is the ability to process large volumes of documents daily, which makes manual analysis unfeasible. Furthermore, the system uses a rules-based robot to pre-classify documents, which are later reviewed by experts, ensuring greater accuracy in classifications. The integration of expert feedback also allows for continuous improvement of the model. Another positive aspect is the possibility of quickly identifying acts that may negatively impact the environment, contributing to greater public awareness and more effective monitoring of government actions. This combination of automation and human supervision makes the *DeepPolicyTracker* tool powerful for environmental protection in Brazil.

Finally, the work [Zhang et al. 2024] introduces a hybrid approach combining large language models (LLMs) and taxonomy enrichment for document classification with minimal supervision. One of its primary strengths is its innovative use of LLMs to annotate core classes, thereby reducing the need for manual supervision. TELEClass also enhances classification accuracy by incorporating semantic features into its hierarchical taxonomy, refining the classification results. This method proves highly efficient for handling large and complex document sets, such as those in government applications, where fully supervised methods may be impractical.

The HALF method leverages NLP and machine learning to classify publications from official documents, distinguishing them through a hybrid human-in-the-loop strategy. Unlike DODFMiner, which focuses on Named Entity Recognition (NER), HALF prioritizes subject classification with iterative dictionary refinement. It aligns with ASKE's semantic similarity calculations but differs by avoiding zero-shot classification. HALF also integrates semi-supervised clustering, similar to [Dobša and Kiers 2022], but exclusively incorporates human feedback for subject refinement. Compared to DeepPolicyTracker, which combines BERT and expert validation, HALF extends beyond a specific policy area. TELEClass shares HALF's goal of minimizing supervision through LLMs, but refines taxonomies, while HALF dynamically updates a subject dictionary. Overall, HALF ensures robust classification by integrating hierarchical similarity analysis, iterative human validation, and subject refinement.

4. Methods

In this section we present the proposed method to identify and classify the subjects of publications present in the Official Gazettes of the State of Ceará [1]. These Gazettes are official periodic publications that document a wide range of information and administrative acts related to state government. They serve as an official means of communication between the government and the public, offering transparency and access to important information about public administration.

Among the methods present in this article is the use of large language models to classify publications from each organization. In addition to the use of LLMs, the study also includes essential data pre-processing steps, such as normalization, tokenization, vectorization, and the use of a word dictionary composed of the most relevant terms

[1] http://pesquisa.doe.seplag.ce.gov.br/doepesquisa/

that appear in each publication.

4.1. Human-Assisted Labeling Feedback Method for Subject Mining

The HALF method proposed follows the following steps: In step one, we download the Official Gazettes of the State of Ceará in PDF format. After that, the content of each PDF is extracted based on the most current structure of the official journals (12/2022 - present), and the data from each journal is stored in separate JSON files and organized by Official Gazette date. We chose this file type due to the compact size and search speed of all content previously contained in PDFs. Each Gazette publication will appear in the corresponding JSON file with the following information: ”DATE”, ”NOTEBOOK”, ”PAGE”, ”ORGAN NAME”, ”TEXT”, and ”HIGHLIGHTS”.

Once we had completed the data extraction and processing, we separated the data into JSON files. In step two, we then created an external dictionary by manually inserting the most relevant words that appear in each publication of the processed journals. These words are mostly words that appear in bold in the publications. This dictionary contains 152 words. The purpose of this dictionary is to make it easier for an LLM to define a subject for each publication.

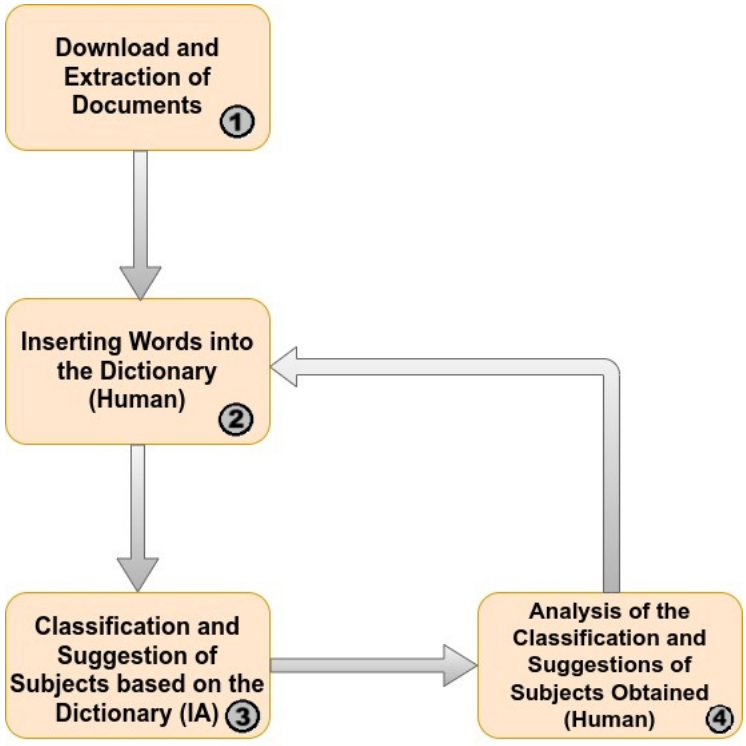

Figure 1. HALF Flowchart

With the dictionary created, in step three we use LLM, in this case GPT-4o Mini, to assign a set of subjects, like ”Portaria”, ”Edital de Convocação” and ”Reconhecimento de Dívida”. The method define the subject of each publication based on a prompt.

In step four, if LLM suggests words that are not present in the dictionary, a human will analyze the suggestions and manually insert them into the dictionary (step two again)

so that in the next iterations LLM can suggest fewer and fewer words as suggestions for publication subjects and the definition of subjects is faster. The flow described from the JSON files generated until the definition of subjects for publications in the Official Gazettes of the State of Ceará by LLM is what defines the Human-Assisted Labeling Feedback Method for Subject Mining (HALF). These steps are described in Figure1.

After defining each publication's subject using GPT-4o Mini via the prompt 4.1, the result is stored in a new "SUBJECT" field, updating the JSON dataset accordingly for all entries.

Your task is to respond only with the subject that best describes the publication;

If there is a valid combination of subjects that best describes the text, join one, two or three subjects, separating them with / as it best fits according to the examples:

Examples = [[portaria/autorização/viagem, portaria/exoneração,portaria/concessão /diárias, autorização/viagem/diárias]

If the text contains a "Portaria", "Decreto", "Corrigenda" or "Lei", it should appear first in the list of results;

If none of them describe it correctly, just return "No Subject Matches:" and a suggestion with 1 to 5 words.

Subjects = ['portaria','decreto','lei','corrigenda','revisão','demissão', 'arrecadação','denominação','acréscimo','reconhecimento','declaração','concessão', 'alteração', 'fixação','Atualização','Criação','divulgação','cessão','aviso', 'concorrência','instrução normativa','instrução','citação por edital', 'edital de intimação','edital de convocação','edital de notificação','edital','convocação', 'ato declaratorio','licenciamento','decreto','portaria administrativa','indenização', 'pensão','reconhecimento de dıvida','reconhecimento de despesa','nomeação', 'valestransportes','arquivamento','reconhecimento','despesa','divida','exoneração', 'doação','registro de preços,'registro','reestruturacao','autorização','designação', 'constituição','composição','relatório','quitação','conclusão','aditivos aos contratos','aditivo ao contrato','aditivo de convênio','contratação','aditivo', ...']

Prompt 4.1. Prompt used in GPT-4o Mini

The task aims to classify publications from the official gazettes of the state of Ceará using a predefined set of subjects. The goal is to identify and return the subject (or a combination of subjects) that best describes the content of the publication. Each publication should be tagged with the most appropriate subject(s). The LLM may select up to three subjects, combining them with a slash ("/") to reflect multiple relevant topics. The combination should make logical and semantic sense based on the content. If the publication includes the words "Portaria", "Decreto", "Corrigenda" or "Lei", this term should be placed at the beginning of the selected subject(s). For example, if a publication refers to a travel authorization and within the text there is the word "Portaria", the three possible sub-

jects would be ”portaria/autorização/diárias”. If none of the available subjects in the set adequately describes the publication, respond with ”No Matching Subject:” followed by a personalized and concise suggestion (1 to 5 words) that captures the essence of the publication’s text. Subjects should be formatted using only lowercase letters and separated by slashes. Examples such as ”portaria/exoneração” or ”autorização/viagem/diárias” serve as models to guide consistent and accurate responses.

5. Results

First, a hierarchical similarity analysis was performed between the 152 subjects of the initial dictionary, manually entered from the analysis of the Official Gazettes of the State of Ceará over a two-year period (December 2022 to December 2024), and the types of subjects available on the SEPLAG website. The SEPLAG website is the page of the Agency that is part of the Government of the State of Ceará, where information and news about this agency are published, including the Official Gazettes of the State of Ceará. The hierarchical approach was necessary to verify whether the SEPLAG subjects, which were mostly highly specific, were covered by the more general subjects of the initial dictionary.

5.1. Evaluation

We generate a ROC (Receiver Operating Characteristic) curve to determine the optimal similarity threshold (which was 63.48 or 0.6348). This threshold defined the point at which a subject from the initial dictionary was considered similar to a subject from SEPLAG. The ROC curve also highlighted the Area Under the Curve (AUC), indicating the effectiveness of the similarity classification like show the Figure 2.

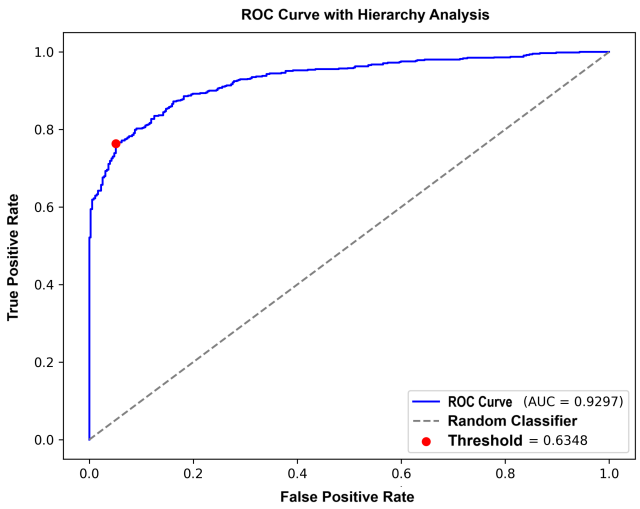

Figure 2. ROC Curve HALF method.

A publications dataset was created, including only those publications associated with similar subjects from both the initial dictionary (152 subjects) and the SEPLAG articles. A total of 1.044 publications were analyzed and categorized according to the following attributes: date, government agency, page, publication number, full text of the

publication, topic assigned by the HALF method, matter type assigned by SEPLAG, and highlighted words (words highlighted within the publication, usually in bold or italics).

Each publication was analyzed using GPT-4o Mini with the prompt and dictionary described in the Methods section. LLM classified each publication after processing its content. A similarity calculation was performed between the subject defined by GPT-4o Mini and the subject types assigned by SEPLAG. This calculation was based on a weighted average of semantic and written similarity, with a higher weight given to semantic similarity (0.7) compared to written similarity (0.3).

The HALF method achieved almost 89% accuracy, meaning that 89% of publications were assigned to subjects similar to those defined by SEPLAG (Figure 3).

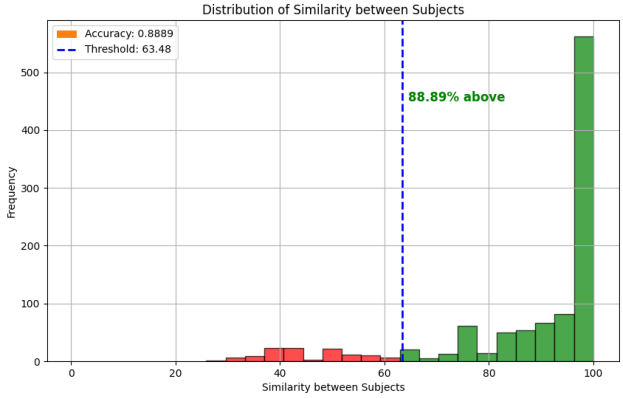

Figure 3. Accuracy HALF method.

We label each publication with 1 or 0, where:

- 1: The subject assigned by the HALF method was similar to the subject assigned by SEPLAG.
- 0: The subject assigned by the HALF method was not similar to the subject assigned by SEPLAG.

Finally, we evaluate the performance with the confusion matrix. It displays the distribution of true positives, false positives, true negatives and false negatives, providing relevant information about the performance of the method (Figure 4). With these two metrics we can comprehensively evaluate the effectiveness of the HALF method in classifying publications, ensuring high accuracy in aligning subjects with those defined by SEPLAG.

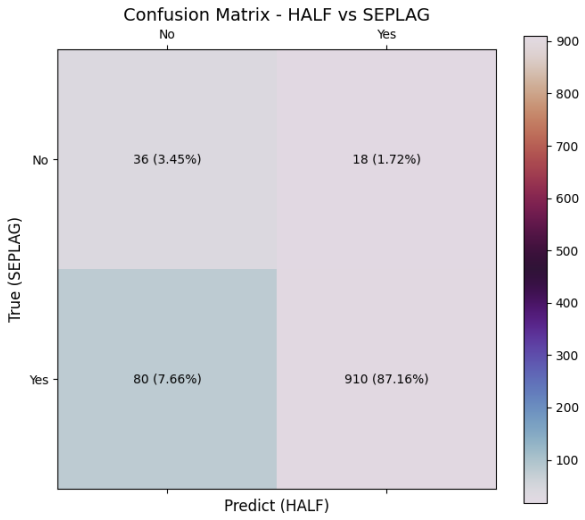

Figure 4. Confusion Matrix HALF method.

5.2. Search Data Mining Tool

By implementing the Human-Assisted Labeling Feedback Method for Subject Mining (HALF) and adding subjects suggested by GPT-4o Mini, we create an online platform with a search tool for each publication based on the content extracted and processed in the JSON files generated from the official gazettes of the State of Ceará.

To use the tool, it is necessary to store the JSON files, which contain several publications, in a directory so that, through JavaScript commands, the tool can filter the result based on the government agency, keywords or even the name of a server and, most importantly, search for a specific publication through the subjects that were defined and inserted in these files through the application of the HALF method, within a period of time specified by the user. Below, in Figure 5, a layout of the tool created to display such information is presented:

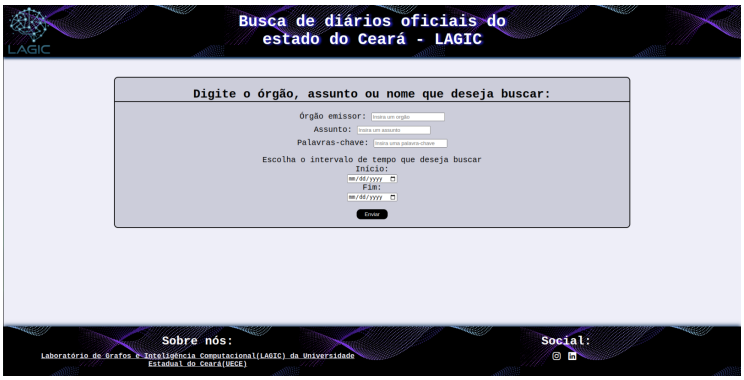

Figure 5. Search Data Mining Tool Layout

5.3. Discussion

The HALF method has demonstrated efficiency in classifying publications with subjects, in some cases classifying publications with more appropriate subjects than those assigned

by SEPLAG. This can be seen in Figure 6. In the image, it is possible to see an example of a publication that was labeled by SEPLAG as ”Contrato Temporário”, but our method, through GPT-4o Mini, classified it as ”Contratação de Professores”. This subject matches the text of the publication. This phenomenon was particularly evident in cases where the subjects predefined by SEPLAG were overly specific or misaligned with the actual content of the publication.

The HALF method, by prioritizing contextual and semantic analysis, was able to generate subjects that better reflected the essence of the publications, as can also be seen from the Figure 6. This highlights the potential of the HALF method not only as a classification tool, but also as a means of refining and improving existing subject taxonomies.

In contrast, we noticed that GPT-4o Mini only suggested one new subject for the publications, even with a relatively small initial subject dictionary (152 words) when compared to the labels used by SEPLAG (2.417 subject types). This suggests that LLM tends to focus on using only the set of words that were passed, instead of proposing new subjects, possibly better labels for the publications.

ÓRGÃO: SECRETARIA DA EDUCAÇÃO
DATA: 16/09/2024
CADERNO: 2
PÁGINA: 28
ASSUNTO SEPLAG: CONTRATO TEMPORÁRIO
ASSUNTO HALF: CONTRATAÇÃO DE PROFESSORES

TEXTO: ”EXTRATO AOS TERMOS DOS CONTRATOS TEMPORÁRIOS DE PROFESSORES - CREDE 8 - BATURITEPROCESSO Nº22001.113488/2024-42 - ADITIVOLOTE 33/2024CONTRATANTE: O Estado do Ceará, através da Secretaria da Educação/ESCOLA: 23054409 - EEMTI DEPUTADO UBIRATAN DINIZ DE AGUIAR. CONTRATADOS: o(s) PROFESSOR(ES): GLORIA MATOS MACIEL - CPF: 90166825387 - MATRÍCULA: 22200181567476 - CARGO: PROF CTPD LIC PLENA - TIPO: HORA-AULA - MOTIVO: LICENCA - MATRÍCULA SUBSTITUÍDO: 22000130372212 - NOME SUBSTITUÍDO: CLEANIA MARTINS DE OLIVEIRA - JUSTIFICATIVA: Licença para Tratamento de Saúde - CRITÉRIO: §1º, ARTIGO 4 - TURNO: I - CH SEMANAL: 3 - CH MENSAL: 15 - VALOR HORA-AULA: R$ 25,63905 - PERÍODO: 31/08/2024 a 29/10/2024 - VALOR MENSAL: R$ 384,59;LUIZIANE ALVES DE LIMA - CPF: 02875937383 - MATRÍCULA: 22200181567174 - CARGO: PROF CTPD LIC PLENA - TIPO: HORA-AULA - MOTIVO: LICENCA - MATRÍ-CULA SUBSTITUÍDO: 22000130372212 - NOME SUBSTITUÍDO: CLEANIA MARTINS DE OLIVEIRA - JUSTIFICATIVA: Licença para Tratamento de Saúde - CRITÉRIO: §1º, ARTIGO 4 - TURNO: I - CH SEMANAL: 2 - CH MENSAL: 10 - VALOR HORA-AULA: R$ 25,63905 - PERÍODO: 31/08/2024 a 29/10/2024 - VALOR MENSAL: R$ 256,39...”

Figure 6. Example of a publication (Translate in Portuguese).

6. Usage Scenario

As discussed previously, the use of AI methods is an advantageous strategy for identifying relevant topics for document classification. The HALF method uses AI methods to classify and present subjects that identify publications in the official gazettes of the state of Ceará.

In this section, we present two scenarios for using the HALF method. The first scenario aims to present an appropriate use for the HALF method and the online platform with a search data mining tool developed for using it to search for texts in official gazettes of the state of Ceará. The other scenario aims to illustrate how the HALF method could be expanded or used as a basis for identifying relevant topics in other types of documents.

The first application of the HALF method is for the online platform we developed. HALF is applied to identify relevant subjects for publications in official gazettes of the state of Ceará, using ChatGPT-4o Mini’s LLM algorithm as a basis. This means it is possible to use our online platform to search for relevant publications for users based on the subjects identified by HALF. This use speeds up the search for important information, such as appointments, information on state financial expenses, employee terminations, vacations, and employee absences and licenses, among others. This information can be used

by individuals interested in a certain subject and by companies that use official journals to obtain information that can be used in their business, such as bidding, for example.

The second application concerns the possibility of adapting the method to different types of documents. HALF was initially created to mine information and identify subjects in a specific group of documents, with its own formatting. However, the adopted strategy was developed in a generalist manner so that it could be applied in different contexts.

After obtaining the documents that will be used to train the LLMs to identify the subjects, the initial separation of the subjects is carried out by a human agent with the support of clustering algorithms, such as K-Means, which allow an initial idea of possible clusters, capable of being used with initial insights to identify these subjects, as well as a non-standard strategy with the help of a human agent that will build an initial dictionary, which will be expanded as interactions using AI algorithms are carried out until a dictionary is considered reasonable by the team using the method proposed in this article. From then on, it is possible to use it in an automated way with the support of an LLM, which in our study was ChatGPT-4o Mini, but which could be another if the team that will use the method believes that it is more attractive to extract the subjects and be used on a platform or through scripts to select the text excerpts that you want to classify in documents.

7. Treats to Validity

Despite the promising results, some threats to validity should be considered. First, the manual creation of the initial dictionary introduces potential bias, as the subject selection depended on human interpretation of the journal publications. Furthermore, the similarity threshold determined by the ROC curve may not generalize well to other datasets, as it was optimized based on the specific characteristics of SEPLAG classifications.

Another limitation is the reliance on GPT-4o Mini, which, despite its advanced natural language processing capabilities, can still generate inconsistent subject assignments in ambiguous cases. To minimize this issue, we evaluated several prompts and compared our results with those of SEPLAG, which is an already consolidated platform. And the results showed that our method classified the documents satisfactorily compared to the classification of this platform.

8. Conclusion

The main objective of this study was to contribute to the advancement of text analysis and classification techniques to provide an efficient solution for managing large volumes of textual data in government or legal contexts.

The HALF method demonstrated high efficiency in classifying official publications with an accuracy of almost 89%, effectively aligning the assigned subjects with those defined by SEPLAG. Furthermore, in some cases, HALF assigned more appropriate subjects than the predefined SEPLAG classifications, suggesting its potential to improve subject categorization in Official Gazettes. Although there are some limitations, the methodology provides a solid basis for automated classification and iterative refinement of subjects.

Future work will involve applying the HALF method to different government or institutional databases to assess its applicability beyond the Official Gazettes of the State

of Ceará. Furthermore, we intend to conduct an in-depth analysis of potential biases in subject classification, such as the implementation of methods to minimize errors. These efforts will promote continued advances in the efficient management and analysis of large volumes of textual data in government and legal environments.

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cação, F. N., Costa, A. R., Unterstell, N., Yonaha, L., Stec, T., and Ishisaki, F. (2021). Deeppolicytracker: Tracking changes in environmental policy in the brazilian federal official gazette with deep learning. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.

Castano, S., Ferrara, A., Furiosi, E., Montanelli, S., Picascia, S., Riva, D., and Stefanetti, C. (2024). Enforcing legal information extraction through context-aware techniques: The aske approach. *Computer Law & Security Review*, 52:105903.

Christopher, D. M., Prabhakar, R., and Hinrich, S. (2008). Introduction to information retrieval.

Dobša, J. and Kiers, H. A. (2022). Improving classification of documents by semi-supervised clustering in a semantic space. In *Conference of the International Federation of Classification Societies*, pages 121–129. Springer International Publishing Cham.

Eisenstein, J. (2018). Natural language processing. *Jacob Eisenstein*, 507.

Guimarães, G. M., da Silva, F. X., Queiroz, A. L., Marcacini, R. M., Faleiros, T. P., Borges, V. R., and Garcia, L. P. (2024). Dodfminer: an automated tool for named entity recognition from official gazettes. *Neurocomputing*, 568:127064.

Jurafsky, D. and Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released January 12, 2025.

Pinto, F. A. D. G., de Barros Santos, J., Lifschitz, S., and Haeusler, E. H. (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. SBC.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Zangari, A., Marcuzzo, M., Rizzo, M., Giudice, L., Albarelli, A., and Gasparetto, A. (2024). Hierarchical text classification and its foundations: A review of current research. *Electronics*, 13(7):1199.

Zhang, Y., Yang, R., Xu, X., Xiao, J., Shen, J., and Han, J. (2024). Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. *arXiv preprint arXiv:2403.00165*.