Classification of Descriptive Sufficiency in Public Expenditure Liquidation Records

Jonathas E. Fujii[1], Gabriel H. T. Moreira[1], Thiago M. Ventura[1], Patícia C. de Souza[1], Raphael S. R. Gomes[1]

Federal University of Mato Grosso, Brazil
jonathasfujii@gmail.com, g.tabordamoreira@gmail.com, thiago@ic.ufmt.br, pathycsouza@gmail.com, raphael@ic.ufmt.br

Abstract. This study proposes a supervised Natural Language Processing approach based on BERTimbau to classify public expenditure liquidation records according to the descriptive sufficiency of their content. The research is grounded on the premise that generic, vague, or overly document-dependent descriptions reduce transparency and hinder institutional and social oversight of public spending. A labeled dataset of 600 records was constructed by random sampling with artificial class balancing from a corpus of 4,083 permanent equipment liquidation records from the State Executive Branch of Mato Grosso, Brazil, for fiscal year 2025, with labeling validated by two independent auditors via Cohen's Kappa ($\kappa = 0.900$). Three supervised strategies were evaluated under stratified 10-fold cross-validation: TF-IDF + SVM (baseline), BERTimbau Feature-Based, and BERTimbau Fine-Tuning. The results demonstrate that both BERTimbau configurations outperformed the baseline with medium-to-large effect sizes, significant for Feature-Based, with Fine-Tuning achieving macro-F1 of 92.66% and accuracy of 92.67%, confirming the feasibility of contextualized Portuguese-language models for supporting preventive, risk-focused governmental auditing.

CCS Concepts: • **Applied computing** → **Computing in government**; • **Computing methodologies** → **Natural language processing**; • **Information systems** → **Information systems applications**.

Keywords: government auditing, bertimbau, text classification, public expenditure, natural language processing

1. INTRODUCTION

The liquidation of public expenditure constitutes a fundamental stage in the budgetary and financial execution cycle, as it is at this point that the creditor's right is verified and the compliance of the delivered goods, provided services, or fulfilled obligations is established [Brasil 1964]. In this context, the textual description of the liquidation record plays a key role in ensuring transparency, traceability, and accountability of public spending, as it should enable a minimum understanding of the executed object, its linkage to the incurred expenditure, and its proper budgetary classification [Fantini 2015; Santos 2016].

However, in administrative practice, these liquidation records are often written in free-form language with substantial heterogeneity, frequently relying on generic and vague expressions such as "invoice payment" or "service provision." This descriptive opacity is directly aligned with the discussion by [Michener et al. 2018] on the selective or "sanitized" nature of active governmental transparency, which is often heavily mediated by public managers in ways that may obscure inconsistencies or limit the scope of institutional and social oversight.

The use of Natural Language Processing (NLP) and supervised Machine Learning techniques has shown promise in supporting the automated analysis of administrative and financial texts in the public

sector. Previous studies have demonstrated the feasibility of automatically classifying public expenditures and governmental records based on traditional textual representations, such as term frequency (TF-IDF), combined with linear or tree-based classifiers, including Support Vector Machines (SVM) and Random Forest [Teixeira et al. 2023; Vale et al. 2023]. Furthermore, deep learning approaches have also achieved robust performance in classifying short, noisy, and weakly structured textual descriptions, such as banking transactions and accounting entries [Kotios et al. 2022; Kotepuchai and Limpiyakorn 2024].

Regarding Portuguese-language modeling, contextualized models such as BERTimbau [Souza et al. 2023; Santana et al. 2022] have stood out for their ability to capture semantic and syntactic nuances of Brazilian Portuguese in text classification and similarity tasks. This capability is particularly relevant for the analysis of public expenditure liquidation records, in which minor lexical variations, synonyms, abbreviations, and domain-specific bureaucratic jargon may correspond to semantically equivalent content.

Unlike traditional thematic classification tasks (e.g., grouping expenditures by budgetary category), assessing descriptive sufficiency requires modeling the intrinsic quality of free-form text to determine its ability to independently convey the actual object of the expenditure. Given that the substantial grammatical heterogeneity of such records renders rigid heuristic approaches (e.g., rule-based methods using regular expressions) ineffective, this study develops a supervised classification method based on BERTimbau, evaluated in both Feature-Based and Fine-Tuning configurations, to screen public expenditure liquidation records according to their descriptive sufficiency, comparing it against a traditional TF-IDF + SVM baseline and providing a methodological foundation for risk-focused automated auditing in the Brazilian public sector.

2. RELATED WORK

The liquidation of public expenditure, regulated by Law No. 4,320/1964, requires that the formal record explicitly specifies the object of the expenditure as a condition for the legal recognition of the obligation by the State. However, the effective quality of this record is structurally conditioned by the discretionary nature of free-text fields. Studies indicate that information proactively disclosed by governments may be *"sanitized" or extensively mediated by public managers* [Michener et al. 2018], making the liquidation record the locus where such mediation materializes at the individual record level. In this sense, descriptive opacity constitutes a structural *feature* of active governmental transparency, whose automated mitigation demands Natural Language Processing approaches.

In the context of Portuguese-language text modeling, BERTimbau [Souza et al. 2023] has established itself as a state-of-the-art monolingual model for Brazilian Portuguese, outperforming multilingual models in textual similarity and sentence classification tasks [Santana et al. 2022]. Studies on text segmentation [Silva et al. 2024a] have shown that models trained on highly specific domains may degrade when applied to different domains, reinforcing the importance of diverse training data. In applications to governmental texts, [Silva et al. 2024b] evaluated domain-adaptive pretraining (DAPT) on BERTimbau and LaBSE, concluding that monolingual models adapted to official gazettes outperform broad multilingual alternatives in capturing complex administrative syntax, and that data quality is more critical than raw volume.

The classification of noisy financial and accounting records shares methodological challenges with the present study. [Kotios et al. 2022] demonstrated that, although traditional term-frequency-based classifiers provide strong baselines, contextualized neural models offer more robust generalization in the presence of orthographic noise and abbreviations. Similarly, [Kotepuchai and Limpiyakorn 2024] showed that contextual interpretation of technical jargon is crucial to mitigating systematic errors. In the Brazilian governmental oversight context, [Vale et al. 2023] classified public procurement objects using TF-IDF + SVM, while [Teixeira et al. 2023] detected accounting manipulations through super-

vised analysis of expenditure-related textual records, showing that municipal managers deliberately reassign formal budgetary codes to conceal excess expenditures, and demonstrating that free-text descriptions contain the factual characterization of expenditures and serve as a ground truth for risk-based auditing.

This study seeks to extend this body of work by investigating the quality and informativeness of liquidation records from a supervised perspective. The identified gap lies in the fact that most studies focus on the thematic classification of expenditures (i.e., "what" is being purchased), whereas this work focuses on descriptive sufficiency (i.e., "how" the expenditure is recorded for legal transparency purposes). Instead of relying solely on sparse frequency-based representations (TF-IDF), this study adopts contextualized models based on BERTimbau, evaluating the performance gains achieved through deep neural representations in the automated screening of records with indications of descriptive insufficiency.

3. METHODOLOGY

The analysis corpus is derived from budgetary and financial execution records for fiscal year 2025, obtained from the financial management system of the State of Mato Grosso, a federative unit of Brazil whose public expenditures fall under the jurisdiction of the State Executive Branch. For this case study, records pertaining to expenditure element 52 (Permanent Equipment and Materials) were selected across all managing units of the State Executive Branch. This element was chosen because it encompasses acquisitions of durable goods, a category in which precise object discrimination is legally required and practically verifiable, making it particularly suitable for evaluating descriptive sufficiency.

To prepare the free-text content for the supervised learning pipeline, reducing irrelevant and encoding-related noise without removing essential structural nuances, a structured preprocessing sequence was applied: missing values and whitespace, character cleaning and minimal normalization and Regex masking.

A labeling guideline was defined based on the legal requirement for explicit specification of the expenditure object established by Federal Law No. 4,320/1964. Each liquidation record was classified into one of two mutually exclusive categories: *Discriminated Object*, when the record independently conveys sufficient information to identify the acquired good or fulfilled obligation; and *Non-Discriminated Object*, when the record contains generic, vague, or tautological descriptions excessively dependent on external documents, precluding minimal autonomous understanding of the liquidated object. Table I presents illustrative examples of the applied criteria.

Table I. Illustrative examples of the labeling criteria

Record	**Label**	**Justification**
Acquisition of 10 swivel chairs for department X	Discriminated	Identifies the acquired good and the purpose of the expenditure.
Payment of Invoice 12345 regarding acquisition	Non-Discriminated	Does not specify the acquired object; requires an external document for comprehension.

A labeled sample of 600 public expenditure liquidation records was constructed by simple random sampling from the original corpus of 4,083 records pertaining to permanent equipment acquisitions across all managing units of the State Executive Branch in 2025. The drawn records were labeled and, once the quota of the majority class was reached, its surplus was discarded, artificially balancing the sample and deliberately over-representing the target class (*Non-Discriminated Object*), which is the primary focus for quality control and screening.

The labeling process involved two independent evaluators, both government auditors. The first evaluator annotated every drawn record; the second reviewed a random 20% subsample (120 of the 600 retained) to validate inter-annotator consistency. Agreement was measured using Cohen's Kappa coefficient (κ), yielding $\kappa = 0.900$ (near-perfect agreement). Disagreements were resolved by joint consensus, producing the final reference dataset. The resulting class distribution is shown in Table II.

Table II. Class distribution of the labeled sample after balancing

Class	**Records**	**Proportion (%)**
Discriminated Object	306	51.0
Non-Discriminated Object	294	49.0
Total	**600**	**100.0**

Three modeling strategies are compared for the binary classification of descriptive insufficiency:

(1) *TF-IDF + SVM* (*baseline*): sparse TF-IDF representations combined with a soft-margin linear SVM, traditionally effective for short administrative texts.
(2) *BERTimbau Feature-Based*: BERTimbau is used solely as a feature extractor; the `[CLS]` token representation (768-dimensional dense vector) is extracted for each record and fed into a linear SVM classifier.
(3) *BERTimbau Fine-Tuning*: full fine-tuning of BERTimbau-Base (`bert-base-portuguese-cased`), pretrained on Brazilian Portuguese over the BrWaC corpus [Souza et al. 2023], with a linear classification layer added on top of the `[CLS]` token representation.

For reproducibility, models were implemented in Python using `scikit-learn` v1.3 and `transformers` v4.35 with `PyTorch`. The source code is publicly available at `https://github.com/jonathasfujii/public-expenditure-descriptive-sufficiency`. Hyperparameter configurations are detailed in Table III.

Table III. Hyperparameter configuration of the experimental protocol

Model	**Hyperparameter**	**Value**
TF-IDF + SVM	n-gram range	$(1, 2)$
	min_df	1
	max_features	3,000
	stopwords	208 (domain)
	C (regularization)	1.0
	kernel	linear
BERTimbau Feature-Based	base model	`bert-base-portuguese-cased`
	pooling	`[CLS]` token
	C (regularization)	1.0
	kernel	linear
BERTimbau Fine-Tuning	learning rate	2×10^{-5} (linear decay)
	epochs	3
	batch size	16
	max sequence length	128 tokens
	optimizer	AdamW
	weight decay	0.01

The task is formulated as binary classification, where each liquidation record h_i is assigned $\hat{y}_i \in \{0, 1\}$, with 0 denoting *Discriminated Object* and 1 denoting *Non-Discriminated Object*. A stratified 10-fold cross-validation protocol was adopted to ensure statistical reliability and preserve the class distribution of the labeled sample across all training and testing splits, so that every fold reproduces the same balanced design.

Reported metrics include accuracy, precision, recall, and macro-F1 score, the primary model se-

lection criterion, as it treats classes equally regardless of their frequency. The aggregated confusion matrix complements the analysis by mapping Type I (false positives) and Type II (false negatives) errors, the latter being the most critical for the effectiveness of public auditing actions.

4. RESULTS AND DISCUSSION

This section presents the experimental results obtained in validating the proposed supervised approach for classifying permanent equipment liquidation records from the State Executive Branch of Mato Grosso for the year 2025. The labeled sample of 600 records is balanced by construction, with 49.0% of records belonging to the *Non-Discriminated Object* class and 51.0% to the *Discriminated Object* class, an artificial distribution that directly informs the interpretation of macro-F1 as the primary evaluation metric. A comparative evaluation of model performance is provided, followed by a class-level breakdown for the BERTimbau-based approaches, the aggregated confusion matrices, computational cost analysis, and statistical significance tests.

4.1 Overall Comparative Model Performance

The stratified 10-fold cross-validation enabled the assessment of the predictive performance of the term-frequency-based baseline and the BERTimbau-based approaches. Table IV summarizes the average performance obtained for each supervised classifier across independent test splits.

Table IV. Performance of the evaluated models in classifying liquidation records

Model	**Accuracy**	**Precision**	**Recall**	**Macro-F1**
TF-IDF + SVM	0.8967 ± 0.0348	0.9005 ± 0.0358	0.8973 ± 0.0346	0.8965 ± 0.0348
BERTimbau Feature-Based	0.9250 ± 0.0271	0.9273 ± 0.0261	0.9253 ± 0.0270	0.9249 ± 0.0272
BERTimbau Fine-Tuning	0.9267 ± 0.0260	0.9289 ± 0.0269	0.9272 ± 0.0260	0.9266 ± 0.0260

Under the primary selection metric (macro-F1), BERTimbau Fine-Tuning achieved the highest mean (0.9266 ± 0.0260), followed closely by BERTimbau Feature-Based (0.9249 ± 0.0272) and the TF-IDF + SVM baseline (0.8965 ± 0.0348). It should be noted, however, that the advantage of Fine-Tuning over Feature-Based is numerically marginal and, as demonstrated in Section 4.5, is not sustained under statistical significance testing. The relevant performance gap therefore lies between the BERTimbau-based approaches and the traditional baseline, not between the two BERTimbau configurations.

4.2 Class-Level Performance of BERTimbau Approaches

To assess detailed discriminative performance under realistic risk-screening scenarios, class-level metrics were computed for both BERTimbau-based strategies. Table V presents these results derived from the aggregated outputs of the cross-validation folds.

Table V. Class-level performance for the BERTimbau-based approaches

Model / Class	**Precision**	**Recall**	**F1-score**
BERTimbau Feature-Based			
Discriminated Object	0.9307	0.9216	0.9261
Non-Discriminated Object	0.9192	0.9286	0.9239
BERTimbau Fine-Tuning			
Discriminated Object	0.9426	0.9118	0.9269
Non-Discriminated Object	0.9112	0.9422	0.9264

Class-level performance is essential to calibrate the practical utility of the model in governmental oversight, as the F1-score for the *Non-Discriminated Object* class directly measures the effectiveness of risk-screening actions. Both BERTimbau configurations present highly balanced metrics across both

classes. The Fine-Tuning model achieved F1-scores of 0.9269 for adequately described records and 0.9264 for risk records, while the Feature-Based model yielded similarly balanced results (0.9261 and 0.9239, respectively).

4.3 Confusion Matrix and Error Distribution

The distribution and nature of predictive errors were assessed through the aggregated confusion matrices from all 10 cross-validation folds, presented in Table VI.

Table VI. Aggregated confusion matrices across 10 folds for all three approaches

Model / Actual Class	Predicted: Non-Disc.	Predicted: Disc.
TF-IDF + SVM		
Non-Discriminated Object	271	23
Discriminated Object	39	267
BERTimbau Feature-Based		
Non-Discriminated Object	273	21
Discriminated Object	24	282
BERTimbau Fine-Tuning		
Non-Discriminated Object	277	17
Discriminated Object	27	279

From the perspective of governmental control and auditing, the confusion matrices reveal two operationally distinct error types. False positives (adequately described records incorrectly flagged as insufficient) represent a moderate increase in redundant manual verification effort. Across models, both BERTimbau configurations reduced these from 39 (TF-IDF + SVM) to 24 and 27, lowering the rate of false administrative alerts. False negatives (records with descriptive insufficiency that pass undetected) represent the critical auditing risk, as they correspond to opaque expenditures that escape automated screening. The Fine-Tuning model achieved the strongest resilience to this error, recording only 17 omissions among the 294 actual positives (5.8%), against 21 (7.1%) for Feature-Based and 23 (7.8%) for the baseline.

The two BERTimbau configurations thus split the error profile: Fine-Tuning yields the fewest false negatives (17 vs. 21) and Feature-Based the fewest false positives (24 vs. 27), with 44 and 45 errors against 62 for the baseline. It must be noted, however, that the difference between the two BERTimbau configurations corresponds to only a few records and falls within normal sampling variation, and should not be interpreted as evidence of configurational superiority, as confirmed by the statistical tests in Section 4.5.

4.4 Computational Cost

To assess the practical viability of deploying each strategy in a governmental control environment, the processing time accumulated across the 10 cross-validation folds was measured for each approach. Table VII details the computational cost and the hardware infrastructure used.

The results reveal a clear trade-off between analytical performance and computational complexity. The TF-IDF + SVM baseline completes the entire pipeline in 0.62 seconds on local CPU, reflecting minimal infrastructure requirements. The BERTimbau Feature-Based approach requires 34.00 seconds for embedding extraction on local CPU, representing nearly the entirety of its processing cost (34.18 seconds total), since the subsequent SVM classifier requires only 0.17 seconds. The BERTimbau Fine-Tuning presents the highest computational demand, requiring 373.05 seconds (approximately 6 minutes and 13 seconds) for full parameter optimization on cloud-based GPU acceleration (Google Colab). For governmental agencies with limited infrastructure, the Feature-Based configuration offers a viable intermediate option, providing contextualized representations without requiring GPU

resources or full model retraining.

Table VII. Accumulated processing time across 10 cross-validation folds

Model	Environment	Embedding Extraction	Training	Total
TF-IDF + SVM	Local CPU	–	0.56 s	0.62 s
BERTimbau Feature-Based	Local CPU	34.00 s	0.17 s	34.18 s
BERTimbau Fine-Tuning	GPU (Colab)	–	368.11 s	373.05 s

4.5 Statistical Significance Analysis

To verify whether the observed differences in macro-F1 across the three experimental strategies reflect genuine performance distinctions or sampling fluctuation, paired hypothesis tests were conducted. The paired Wilcoxon signed-rank test was adopted as the primary method, recommended for small sample sizes ($N = 10$ folds), complemented by the paired Student's t-test as a parametric reference. Effect sizes were measured through Rank-Biserial Correlation (RBC) for the non-parametric test and Cohen's d for the parametric approach. The null hypothesis of performance equivalence was tested at $\alpha = 0.05$. Results are reported in Table VIII.

Table VIII. Paired statistical test results across folds ($N = 10$)

Paired Comparison	Wilcoxon (p)	RBC	t-test (p)	Cohen's d
BERTimbau FT vs. TF-IDF + SVM	0.0645	0.6909	0.0412	0.7529
BERTimbau FT vs. BERTimbau FB	0.7695	0.1455	0.8758	0.0508
BERTimbau FB vs. TF-IDF + SVM	0.0382	0.7778	0.0415	0.7516

At the 5% significance level, the comparison between BERTimbau Feature-Based and the TF-IDF + SVM baseline yielded statistically significant results in both tests ($p = 0.0382$ for Wilcoxon, $p = 0.0415$ for Student's t-test), supported by medium-to-large effects (RBC = 0.7778, Cohen's d = 0.7516). In the comparison between BERTimbau Fine-Tuning and the baseline, the parametric test indicated significance ($p = 0.0412$, Cohen's $d = 0.7529$), while the Wilcoxon test pointed to a marginal tendency ($p = 0.0645$, RBC = 0.6909), reflecting the distributional characteristics of the fold-level scores with $N = 10$ degrees of freedom. Finally, the direct comparison between Fine-Tuning and Feature-Based showed no statistically significant difference ($p = 0.7695$ for Wilcoxon, $p = 0.8758$ for t-test, Cohen's $d = 0.0508$), confirming that both BERTimbau configurations deliver equivalent performance under the tested conditions.

Two caveats apply. The folds are not independent, as their training sets overlap by construction, and paired tests over them behave liberally, inflating significance and effect sizes [Dietterich 1998]; nor was any correction for multiple testing applied, under which (Bonferroni, $\alpha = 0.0167$) no observed p-value would remain significant. The gains over the baseline are therefore consistent evidence rather than a formal claim.

5. CONCLUSION

This study demonstrated the feasibility of using supervised NLP models to classify public expenditure liquidation records according to their descriptive sufficiency, addressing a gap in governmental oversight research that has focused primarily on thematic expenditure classification rather than the intrinsic quality of administrative records. The experimental results confirm that BERTimbau-based approaches outperform the TF-IDF + SVM baseline with medium-to-large effect sizes, significant for the Feature-Based setting, answering the research question posed. Notably, the equivalence between the Fine-Tuning and Feature-Based configurations, combined with the substantially lower computational cost of the latter, positions BERTimbau Feature-Based as the most operationally viable option for continuous auditing workflows in resource-constrained governmental environments.

Three limitations of this study must be acknowledged. The corpus is restricted to element 52 for fiscal year 2025 in the State Executive Branch of Mato Grosso, and results should not be generalized to other elements, years, or entities without further validation. Furthermore, class balance was obtained artificially, by discarding surplus majority-class records, so the reported metrics refer to a balanced setting rather than to the actual prevalence of insufficient records in the full liquidation universe. The paired tests also rely on overlapping folds and no correction for multiple comparisons, so the p-values are indicative.

As future work, the following directions are recommended: (i) expanding the labeled dataset to cover multiple expenditure elements and government entities, improving classifier generalization; (ii) integrating Active Learning techniques to reduce annotation costs by prioritizing borderline and high-uncertainty cases; (iii) incorporating explainability methods such as LIME or Integrated Gradients to provide auditors with interpretable justifications for predictions of descriptive insufficiency; and (iv) integrating the classifier into the continuous data pipeline of the State Comptroller's Office, enabling preventive alerts focused on descriptive integrity risks.

ACKNOWLEDGMENT

The authors declare that Claude (Anthropic) was used for the analysis of conceptual ideas and for grammatical review of written texts in English. All content was critically reviewed and validated by the authors.

REFERENCES

Brasil. Lei n.º 4.320, de 17 de março de 1964. estatui normas gerais de direito financeiro para elaboração e controle dos orçamentos e balanços da união, dos estados, dos municípios e do distrito federal, 1964.

Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10 (7): 1895–1923, 1998.

Fantini, W. d. S. *Publicação de Dados Conectados sobre Despesas Orçamentárias do Governo Federal Brasileiro.* M.S. thesis, Centro de Informática, Universidade Federal de Pernambuco, Recife, PE, Brasil, 2015.

Kotepuchai, P. and Limpiyakorn, Y. Tree-based classifiers for smart general ledger code suggestion. In *Proceedings of the 2024 13th International Conference on Informatics, Environment, Energy and Applications (IEEA 2024).* ACM, Tokyo, Japan, pp. 17–22, 2024.

Kotios, D., Makridis, G., Fatouros, G., and Kyriazis, D. Deep learning enhancing banking services: a hybrid transaction classification and cash flow prediction approach. *Journal of Big Data* 9 (1): 130, Oct., 2022.

Michener, G., Contreras, E., and Niskier, I. Da opacidade à transparência? avaliando a Lei de Acesso à Informação no Brasil cinco anos depois. *Revista de Administração Pública*, 2018. RAP, Rio de Janeiro. Recebido set. 2017; aceito abr. 2018. [versão traduzida].

Santana, I. N., Oliveira, R. S., and Nascimento, E. G. S. Text classification of news using transformer-based models for portuguese. *Journal of Systemics, Cybernetics and Informatics* 20 (5): 33–59, 2022.

Santos, L. F. d. *Aplicação de Machine Learning na Classificação de Restos a Pagar.* M.S. thesis, Escola de Administração de Empresas de São Paulo, Fundação Getulio Vargas, São Paulo, SP, Brasil, 2016.

Silva, L. A. C., Rodrigues, M. S. F., Archanjo, A. P., Pessoa, L., Silva, M. L., Almeida, T. F., and Silveira, L. Segmentação textual baseada em tópicos em português utilizando bertimbau. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL).* SBC, Porto Alegre, RS, Brasil, pp. 32–36, 2024a.

Silva, M. O., Oliveira, G. P., Costa, L. G. L., and Pappa, G. L. Evaluating domain-adapted language models for governmental text classification tasks in Portuguese. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados (SBBD 2024).* Sociedade Brasileira de Computação, Florianópolis, SC, Brasil, pp. 247–259, 2024b.

Souza, F. C., Nogueira, R. F., and Lotufo, R. A. BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. *Applied Soft Computing* vol. 149, pp. 110901, 2023.

Teixeira, P. H., da Silva, N. F. F., and Salvini, R. Machine learning based method for auditing personnel expenses in public expenditure. In *Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC).* IEEE, Torino, Italy, pp. 320–327, 2023.

Vale, A. H., Santos, P., Soares, H., and Moura, R. S. Automatic classification of public expenses in the fight against COVID-19: A case study of TCE/PI. In *Proceedings of the XIX Brazilian Symposium on Information Systems (SBSI 2023).* ACM, Maceió, AL, Brasil, pp. 118–129, 2023.