Machine Learning-based Classification of Portuguese News Articles on Public Procurement Fraud

Paulo Marcos de Assis[1], **Márcio Castro**[1], **Jônata Tyska Carvalho**[1]

[1]Department of Informatics and Statistics – Graduate Program in Computer Science
Federal University of Santa Catarina (UFSC)

`paulo.marcos@grad.ufsc.br, {marcio.castro, jonata.tyska}@ufsc.br`

Abstract. *Combating fraud in public procurement is a critical task for oversight agencies, and the news media offer a rich source to uncover irregularities. Yet, the sheer volume of daily news hampers the identification of relevant reports. We propose a text classification pipeline to flag news articles reporting procurement fraud. Contributions include a labeled dataset of 9,412 news articles, the optimization and evaluation of multiple machine learning methods, and the identification of an effective feature extractor-classifier combination. Using BERT embeddings and Support Vector Machines, we achieved an F1-Score of 100% on a test set of 8553 news articles with only 0.15% positive labels. Results offer a promising news-based methodology for evidence-based fraud detection.*

1. Introduction

Effective oversight of public administration depends on investigative and monitoring authorities obtaining timely and accurate information. In public procurement, news coverage can be an essential data source to uncover potential fraud and initiate investigations [de Souza and Dorneles 2025]. However, prosecutors face significant information overload from the massive daily news production. For instance, the two leading portals in the Brazilian state of Santa Catarina [1] ("NSC Total" and "ND Mais") publish together over 500 articles daily.

To address this gap, this work develops and evaluates a text classification pipeline to automatically identify news articles reporting fraud in public tenders. Primary contributions include: (1) a labeled dataset of 9,412 Portuguese news articles from Santa Catarina portals, developed through the "Céos" project [2] — a partnership between Santa Catarina's Public Prosecutor's Office and Federal University to develop AI-based fraud detection tools; and (2) we optimize and evaluate different Machine Learning (ML) methods for automated classification, testing Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) algorithms. To identify the most effective combination of feature extractor and classifier, we compared two variants of BERTimbau [Souza et al. 2020] (base and large) against the Term Frequency-Inverse Document Frequency (TF-IDF). The BERTimbau-Large+SVM model achieved perfect performance (100% precision and recall) tested on 8,553 unseen web-scraped articles from NSC Total, correctly identifying all 13 public procurement fraud cases.

2. Related Work

Combating public procurement fraud is a critical concern for public oversight bodies. In this sense, *Nai et al.* highlight the growing adoption of Artificial Intelligence (AI),

[1] `https://www.nsctotal.com.br` and `https://ndmais.com.br`
[2] `https://ceos.ufsc.br/`

particularly Machine Learning and Natural Language Processing (NLP), for analyzing procurement data and identifying irregularities. Text-based fraud has recently gained attention [Nai et al. 2022]. *Lima et al.* used BERTimbau — a Brazilian Portuguese variant of BERT — to extract fraud indicators from official procurement documents, achieving an F1-score of 0.86 and outperforming baselines [Lima et al. 2023]. However, their approach remains limited to formal procurement documents, while the news media remain largely underexplored. Our work seeks to fill this gap by applying NLP techniques to news articles as a complementary data source for public procurement monitoring. To our knowledge, no previous work has proposed a Portuguese-language approach to detect potential procurement fraud through journalistic content.

3. Methodology

This section describes the dataset constructed for this study and the methodology used to develop and evaluate the automated news classification pipeline. Figure 1 illustrates the end-to-end pipeline.

First, we constructed a gold-standard labeled dataset with news articles annotated as positive (“1”) for procurement fraud cases or negative (“0”) for other topics. Our training set has 859 labeled news stories (342 positive, 517 negative) from 112 different news sources from Santa Catarina. The test set comprises 8,553 unseen news articles from NSC Total (March-June 2025). Manual labeling of the entire test set identified only 13 cases related to public procurement fraud, a severe class imbalance (0.15% positive cases), reflecting the rarity of the topic in daily news coverage. Table 1 presents our training and test sets.

Next, we compared traditional TF-IDF vectorization and transformer-based embeddings [Vaswani et al. 2017] for feature extraction. While TF-IDF assigns weights based on term frequency without semantic context, news texts often use varied terminology like “public contract favoritism” or “contracting irregularities” (and others) rather than explicit “procurement fraud” mentions. To evaluate the effectiveness of contextual understanding, we implement transformer-based embeddings using BERT [Devlin et al. 2019] for its proven ability to capture semantic context. For our Brazilian Portuguese corpus, we adopted BERTimbau [Souza et al. 2020], testing both base (110M parameters) and large (335M parameters) models as feature extractors for SVM, LR, and RF classifiers.

Finally, we evaluated nine different model configurations that combine three classifiers (LR, SVM, and RF) with three feature extraction methods: TF-IDF, BERTimbau-base, and BERTimbau-Large. All models were trained on the 859 labeled dataset, with performance evaluated using 5-fold cross-validation, and all classifiers were optimized via GridSearchCV, with hyperparameters tuned separately for each feature extractor. Our ex-

Table 1. Main characteristics of the labeled dataset

Feature	Training	Test	Total
Total articles	859	8553	9412
Number of news portals	112	1	112
Positive class (procurement fraud)	342	13	355
Negative class	517	8540	9057
Average article length (words)	491	524	521
Deviation length (words)	379	289	299
Min length (words)	14	102	14
Max length (words)	4946	4569	4946
Collection period	04/2020 - 06/2025	03/2025 - 06/2025	04/2020 - 06/2025

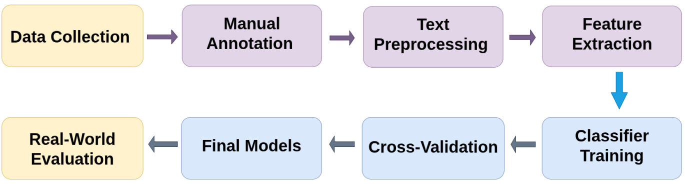

Figure 1. End-to-end pipeline for news classification, from data collection and annotation to training, validation, and evaluation.

periments used Python with Scikit-Learn. After text normalization, BERTimbau embeddings were generated via `SentenceTransformer` library with mean pooling, producing 768-dimensional feature vectors (base model) and 1024-dimensional vectors (large model). The cross-validation results (see Table 2) showed that LR performed best with the TF-IDF features, SVM dominated with BERT embeddings and RF yielded slightly lower, but consistent scores. Performance metrics emphasized recall given the critical importance of not missing potential fraud cases in oversight applications.

Table 2. Cross-Validation results during training (859 labeled set)

Feature Extractor	Algorithm	Accuracy	Precision	Recall	F1-Score
TF-IDF	**L. Regression**	**0.940**	**0.945**	**0.903**	**0.923**
	SVM	0.932	0.927	0.900	0.913
	Random Forest	0.915	0.889	0.897	0.893
BERT base	**L. Regression**	0.880	0.838	0.865	0.851
	SVM	0.892	0.861	0.871	0.866
	Random Forest	0.887	0.896	0.809	0.851
BERT Large	**L. Regression**	0.894	0.859	0.877	0.868
	SVM	0.897	0.867	0.877	0.872
	Random Forest	0.902	0.913	0.833	0.871

4. Results and Discussion

Given that all three classifiers demonstrated very similar performance during cross-validation, we tested all models on the 8,553 unseen news articles (see Table 3), rather than selecting only the best performing model. This real-world evaluation on a highly unbalanced dataset (0.15% positive cases) revealed critical performance disparities, highlighting the model's varying abilities to handle class imbalance and the importance of testing beyond controlled datasets. Although Logistic Regression achieved acceptable results with the TF-IDF features (F1=0.88), it failed drastically with BERT embeddings, generating 69 and 97 false positives. Random Forest showed poor generalization with the TF-IDF features (F1=0.50), but strong performance with the BERT variants (F1=0.88 and 0.96). SVM demonstrated consistent superiority with BERT-Large (F1=1.0), correctly identifying all procurement fraud articles with zero false positives and zero false negatives.

These contrasting results indicate that SVM is best suited to handle dense, high-dimensional contextual embeddings produced by BERT, demonstrating superior generalization on severely imbalanced datasets. Logistic Regression performed optimally with sparse, traditional feature representations, while RF improved with richer embeddings. This confirms how classifier choices fundamentally depend on feature characteristics.

The perfect score should be interpreted with parsimony. A possible data leakage and the limited size and news source of the test set may have influenced this result. For a more reliable generalizability estimate, future work using larger and more diverse datasets is necessary.

Table 3. Real-world test - 8,553 labeled news articles (only 13 positive cases)

Feature Extractor	Algorithm	TP	FP	TN	FN	Precision	Recall	F1-score
TF-IDF	L. Regression	11	1	8539	2	0.916	0.846	0.880
	SVM	12	3	8537	1	0.800	0.923	0.857
	Random Forest	9	14	8526	4	0.391	0.692	0.500
BERT (base)	L. Regression	13	69	8471	0	0.158	1.000	0.273
	SVM	13	2	8538	0	0.867	1.000	0.928
	Random Forest	11	1	8539	2	0.916	0.846	0.880
BERT (Large)	L. Regression	13	97	8443	0	0.118	1.000	0.211
	SVM	13	0	8540	0	**1.000**	**1.000**	**1.000**
	Random Forest	12	0	8540	1	1.000	0.923	0.960

5. Conclusion

This work presents an automated pipeline for classifying news articles on procurement fraud, contributing with a gold standard dataset of 9,412 labeled news articles. In addition, the perfect performance of BERTimbau-Large+SVM demonstrates the substantial advantage of contextual language understanding to detect procurement fraud in news articles. This performance suggests a strong potential for practical deployment in oversight applications, where potential reports on unidentified bidding fraud cases can have serious consequences for public resources. Future work includes the implementation of GPT-based models and the investigation of classification performance using different LLMs. We also plan to expand the current dataset with additional news articles, enhancing the training capacity of our models.

6. Acknowledgements

This work was funded by the Ministério Público de Santa Catarina (MPSC)/Brazil through the Project Ceos: Data Intelligence for the Society.

References

de Souza, A. C. S. and Dorneles, C. F. (2025). Cono: Um coletor automatizado de notícias sobre corrupção em santa catarina. In *Anais da XX Escola Regional de Banco de Dados (ERBD)*, pages 129–132, Florianópolis/SC. Sociedade Brasileira de Computação.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lima, W., Lira, R., Paiva, A., Silva, J., and Silva, V. (2023). Methodology for automatic extraction of red flags in public procurement. In *International Joint Conference on Neural Networks (IJCNN)*, pages 01–07, Gold Coast, Australia.

Nai, R., Sulis, E., Meo, R., et al. (2022). Public procurement fraud detection and artificial intelligence techniques: a literature review. In *International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 1–13. CEUR-WS.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. *Intelligent Systems*, pages 403–417.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008.