Geoparsing of Spatial Coordinates in Brazilian Land Tenure Documents with Deep Learning

Taís Pereira[1,*], Mauro Alixandrini[1], Jorge Pedreira Junior[1] and Vivian Fernandes[1]

[1]Department of Transportation Engineering and Geodesy, Federal University of Bahia (UFBA), Salvador, Brazil

Abstract

This study proposes a modular pipeline to recognize and extract spatial coordinates from unstructured Brazilian land-related documents, expanding geoparsing beyond the dominant focus on toponyms in English-language corpora. The workflow comprises OCR of PDF documents, WordPiece tokenization using the BERTimbau tokenizer, manual annotation of coordinates in Label Studio, dataset construction in JSONL, fine-tuning of a BERTimbau-based sequence labeling model with a CRF layer, and evaluation with seqeval-based metrics. A pilot experiment with five documents and document-grouped cross-validation indicates that the pipeline is functional under very limited supervision, achieving a mean entity-focused F1 around 0.51 for the best hyperparameter setting. The findings provide an initial baseline and motivate future expansion to larger and more heterogeneous corpora.

Keywords

Geoparsing, Named Entity Recognition, BERTimbau, Spatial Coordinates

1. Introduction

Geoparsing is the process of identifying location references, such as toponyms, addresses, and spatial relations, in unstructured text and enriching them with spatial coordinates [13, 7]. It is an active research challenge that has been extensively investigated over the past two decades [8, 9, 14, 17] through rule-based, statistical learning–based, and gazetteer matching–based approaches. Recently, integration of deep learning models, such as BERT and its variants, into statistical learning frameworks (often framed as a subtask of Named Entity Recognition (NER)) has yielded significant improvements in recognition accuracy [21, 16, 10, 4], establishing a promising research frontier within the field.

Despite scientific advances, geoparsing still faces relevant research gaps, including the scarcity of studies addressing multilingual settings: most research focuses on English-language texts, largely because a substantial portion of the available research literature is published in English [6]. Moreover, another relatively underexplored gap with strong practical potential in boundary-description contexts is the development of mechanisms to recognize spatial coordinates in text. In many countries, land administration and land tenure regularization rely on documents that contain spatial descriptions of land parcels; therefore, processing such documents quickly and accurately is highly relevant, as it can reduce errors that lead to land disputes, optimize the analysis of large volumes of textual data, and mitigate delays in procedures related to permitting development projects.

Contributions. Part of an ongoing master's thesis research, this study proposes using the BERTimbau model [19], pre-trained in large-scale Brazilian Portuguese corpora, to extract spatial coordinates from unstructured land-related documents. This choice provides better linguistic representation than multilingual models and has shown strong performance in Portuguese NER tasks [1]. Unlike approaches focused on English toponyms, this study addresses a gap by extracting coordinates from non-English texts, supporting applications in engineering, urban planning, cadastre, and land tenure regularization.

GeoExT 2026: Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026, April 2, 2026, Delft, The Netherlands

*Corresponding author.

tsp.sousapereira@gmail.com (T. Pereira); mauro.alixandrini@ufba.br (M. Alixandrini); jorge.ubirajara@ufba.br (J. P. Junior); vivian.fernandes@ufba.br (V. Fernandes)

2. Methodology

The pipeline developed for this research, illustrated in Figure 1, prioritizes low-cost replication and scientific transparency. By leveraging Google Colab for cost-effective access to high-performance GPUs, the methodology ensures feasibility for academic environments. Furthermore, the workflow adopts a modular design in which each stage generates intermediate artifacts that serve as inputs for subsequent phases. This decoupled structure not only facilitates systematic verification and localized error correction without requiring the entire process to be re-executed, but also enhances traceability. To support full reproducibility, the code and trained models will be made available in a public GitHub repository upon completion of the research.

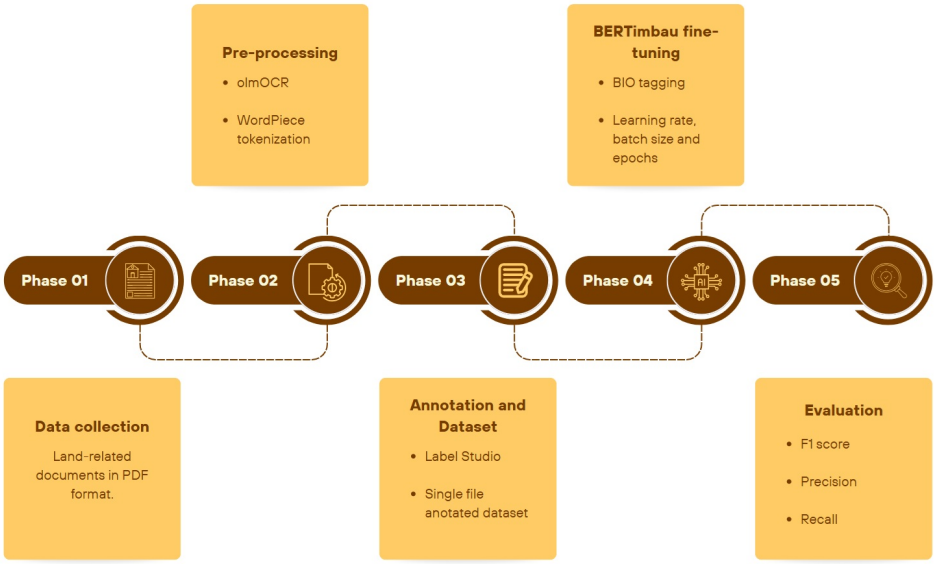

Figure 1: Research pipeline.

2.1. Data collection

The data used in this study comprise official documents issued by a Brazilian institution responsible for conferring legal validity on real estate transactions (2º Registro de Imóveis de Salvador-BA[1]). These documents describe boundary polygons corresponding to land parcels and include spatial coordinates of their vertices, expressed as geographic (DMS or decimal degrees) or projected (UTM) systems. The pipeline focuses on recognizing coordinate expressions as textual entities, independent of format or coordinate reference system. This information can be integrated into GIS workflows. For the pipeline pilot test, five PDF documents were selected, each containing at least 20 vertices.

2.2. Pre-processing

To ensure data consistency, pre-processing began with OCR using the olmOCR-2-7B-1025 tool (Allen Institute for AI[2]) on a Google Colab A100 GPU, essential for processing land documents in scanned PDF format. Subsequently, text was tokenized via the WordPiece method using the BERTimbau Base model tokenizer [19] to leverage its knowledge acquired during pre-training. The process culminated in a structured .csv file (including metadata such as id, page, block, and n_tokens) for Label Studio import. This structured output ensures full traceability between the raw text and the segments used during fine-tuning.

[1] https://www.2risalvador.com.br/
[2] https://olmocr.allenai.org/

2.3. Annotation and Dataset Construction

The author conducted the data label manually through Label Studio[3], ensuring a precise identification of the coordinate spans. Three types of entities (NOR_UTM, LES_UTM, and GEO_UTM) were annotated, totaling 649 entity spans. Individual JSON exports were then consolidated into a single JSONL file to standardize structures and ensure alignment between token and labels. The dataset includes tokens, labels and metadata (e.g., id, doc_id, and entities) to support process traceability and quality control.

2.4. BERTimbau fine-tuning

To optimize performance and manage the computational costs of Transformer-based adaptation, BERTimbau was fine-tuned on high-capacity GPU (A100) within Google Colab. Training was formulated as a token-level classification task using the BIO (Begin, Inside, Outside) tagging scheme. This setup leveraged the model's pre-trained linguistic knowledge to refine its recognition of spatial coordinate patterns, while the high-performance hardware prevented memory limitations and reduced training time across experimental configurations (e.g., batch size and sequence length).

Hyperparameters were defined through empirical experimentation and stratified GroupKFold cross-validation, using doc_id (k = 5) as the grouping variable to prevent information leakage between training and validation splits in a small-dataset scenario. A reduced grid search explored learning rates (2×10^{-5}, 1×10^{-5}), batch size (8), and epochs (10, 15), with values informed by similar BERT-based NER literature (Table 1). The model utilized a hybrid BERT+CRF architecture with a class-weighted cross-entropy loss to handle class imbalance. Early stopping and automatic best checkpoint selection were guided by an entity-focused F1-score (entity_f1), excluding the "O" class to ensure sensitivity to minority classes.

Table 1
Hyperparameters of BERT-based models in NER tasks.

Reference	Model	Hyperparameters		
		Learning rate	Batch size	Epochs
Abilio et al. (2024) [1]	BERTimbau e mBERT	5×10^{-5}	16	2
Benali et al. (2022) [3]	BERT-BGRU	1×10^{-4}	16	5
Botero-Aguirre and García-Rivera (2025) [5]	BERT-base for Spanish (BETO)	2×10^{-5}	16	5
Licari and Comande (2024) [11]	ITALIAN-LEGAL-BERT-FP	5×10^{-5}	10	4
	ITALIAN-LEGAL-BERT-SC	2×10^{-5}	18	4
Löffler et al. (2025) [12]	Transformer-based models	3×10^{-5}	32	< 50
Shingleton and Basiri (2024) [16]	Topo-BERT	2×10^{-6}	4	20
Wu et al. (2024) [20]	BERT-base for Chinese/English	1×10^{-3}	8	30

As final outputs of this stage, token-level confusion matrices were generated (with and without the "O" class), together with metrics files reporting the experimental results and the final BERTimbau model trained under the best-performing configuration, including the artifacts required for reproducibility (weights, encoder, tokenizer, and metadata).

2.5. Evaluation of the trained model

Model performance in the pilot test was evaluated using standard metrics: precision, recall, and the F1-score [1, 2, 15, 19]. Precision and Recall were employed to assess prediction exactness and the model's ability to retrieve all relevant entities, respectively. The F1-score served as a harmonic mean of precision and recall, providing a balanced and comprehensive assessment of the model's recognition performance [2, 18].

[3] https://labelstud.io/

3. Experimental Results

The results of the pipeline pilot test using a small annotated dataset are presented in Table 2. The best-performing configuration (lr = 1×10^{-5}, batch size = 8, 15 epochs) achieved a mean_entity_f1 of 0.5151, with low dispersion across folds (std = 0.0174), suggesting moderate and consistent performance in token-level entity identification. This metric excludes the "O" class and penalizes misclassifications among entity labels. In contrast, the seqeval-based metrics yielded lower mean values (F1 = 0.1504, precision = 0.1139, recall = 0.2285), which is consistent with the fact that seqeval requires greater consistency in the formation of complete entities (BIO tag boundaries) and is therefore more sensitive to fragmentation and boundary errors. The computational cost of the best experiment was 92.30 s on average per fold.

Table 2
Evaluation metrics for the proposed methodological pipeline pilot test.

Best hyperparameter setting	Entity F1 (mean)	Entity F1 (std)	F1 score (mean)	Precision (mean)	Recall (mean)
Lr = 1×10^{-5}, Batch size = 8, 15 epochs	0,5151	0,0174	0,1504	0,1139	0,2285

4. Discussion And Outlook

Experiments demonstrate that the proposed pipeline is functional and technically consistent under extremely limited supervised data. In document-level cross-validation (GroupKFold using doc_id as the grouping variable), the best configuration achieved an entity F1-score mean of approximately 0.51, indicating moderate and relatively stable performance of the model. The observed discrepancy between this result and the lower seqeval-based F1-score is attributable to the limitations inherent to a small dataset, where sparse entity occurrences, especially in the validation document, can lead to zero-valued metrics in certain folds. Nevertheless, the model consistently captures entity-related signals with a balanced trade-off between precision and recall, though it still exhibits BIO boundary errors and class confusion—behavior consistent with the current data scale and the dominance of the "O" class.

For the next stages of this research, the pilot test provides a clear direction: expand the training set with additional documents and a wider variety of spatial coordinate formats, in order to mitigate the "document-specific" effect observed under doc_id-based validation. In parallel, we recommend maintaining diagnostic analysis using metrics aligned with the objective (mean entity F1 score) and through systematic error inspection, including the rate of "erased" entities (predicted as "O") and confusion matrices computed without the "O" class, to identify patterns of confusion across entity types and BIO boundaries. In this way, the current results serve as a reliable initial baseline: they demonstrate that the pipeline works, quantify limitations expected given the dataset size, and define concrete improvement targets for the final version of the study.

As a continuation of this work, provided that BERTimbau demonstrates strong performance when trained on a larger dataset, we intend to apply the final model for spatial coordinate recognition to other document types containing coordinates in different writing patterns and textual layouts, such as maps (titles, legends, coordinate tables, and cartographic textual elements), technical reports, and scientific articles. This procedure will enable an external comparative evaluation of performance, allowing us to assess the model's robustness to domain variation, text extraction quality, and notation conventions. The results are expected to support the expansion of the annotated dataset and the adaptation of the pipeline to more heterogeneous document scenarios.

Acknowledgements

This study was conducted as part of Taís Pereira's master's thesis, funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) of the Brazilian Ministry of Education, and supported by the Geospatial Information Research Group (GPGeo) at the Polytechnic School of the Federal University of Bahia (UFBA).

Declaration On Generative AI

The authors are not native English speakers; therefore, ChatGPT was used to assist with English-language editing (grammar correction, spelling, and improving clarity and conciseness) and with the development of Python code. The authors subsequently reviewed the manuscript and the code, made any necessary revisions, and they take full responsibility for the final content.

References

[1] ABILIO, R.; COELHO, G. P.; SILVA, A. E. A. Evaluating named entity recognition: A comparative analysis of mono- and multilingual transformer models on a novel Brazilian corporate earnings call transcripts dataset. Applied Soft Computing, v. 166, 112158, 2024. https://doi.org/10.1016/j.asoc.2024.112158.

[2] BARBON, R. S.; AKABANE, A. T. Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. Sensors, v. 22, n. 21, 8184, 2022. https://doi.org/10.3390/s22218184.

[3] BENALI, B. A.; MIHI, S.; LAACHFOUBI, N.; MLOUK, A. A. Arabic Named Entity Recognition in Arabic Tweets Using BERT-based Models. Procedia Computer Science, v. 203, p. 733-738, 2022. https://doi.org/10.1016/j.procs.2022.07.109 .

[4] BERRAGAN, C.; SINGLETON, A.; CALAFIORE, A.; MORLEY, J. Mapping cognitive place associations within the United Kingdom through online discussion on Reddit. Transactions of the Institute of British Geographers, v. 49, n. 1, p. 1-18, 2024. https://doi.org/10.1111/tran.12669.

[5] BOTERO-AGUIRRE, J. P.; GARCÍA-RIVERA, M. A. Natural Language Processing for Enhanced Clinical Decision Support in Allergy Verification for Medication Prescriptions. Mayo Clinic Proceedings: Digital Health, v. 3, n. 3, 100244, 2025. https://doi.org/10.1016/j.mcpdig.2025.100244.

[6] CHEN, X.; GELERNTER, J.; ZHANG, H.; LIU, J. Multi-lingual geoparsing based on machine translation. Future Generation Computer Systems, v. 96, p. 667-677, 2019. http://dx.doi.org/10.1016/j.future.2017.07.057.

[7] HU, X., ZHOU, Z., LI, H., HU, T., GU, F., KERSTEN, J., FAN, H., KLAN, F. Location Reference Recognition from Texts: A Survey and Comparison. ACM Computing Surveys, v. 56, n. 5, 112, 2023. https://doi.org/10.1145/3625819.

[8] JONES, C. B.; PURVES, R.; RUAS, A.; SANDERSON, M.; SESTER, M.; KREVELD, M. V.; WEIBEL, R. Spatial information retrieval and geographical ontologies: an overview of the SPIRIT project. In: INTERNATIONAL SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 25., 2002, Tampere. Anais [...]. Nova York: Association for Computing Machinery, 2002. p. 387-388. https://doi.org/10.1145/564376.564457.

[9] LEIDNER, J. L.; LIEBERMAN, M. D. Detecting geographical references in the form of place names and associated spatial natural language, SIGSPATIAL Special, v. 3, n. 1, p. 5-11, 2011. https://doi.org/10.1145/2047296.2047298.

[10] LI, X.; ZHANG, W.; WANG, Y.; TAN, Y.; XIA, J. Spatio-temporal information extraction and geoparsing for public Chinese resumes. ISPRS International Journal of Geo-Information, v. 12, n. 9, 377, 2023. https://doi.org/10.3390/ijgi12090377.

[11] LICARI, D.; COMANDÈ, G. ITALIAN-LEGAL-BERT models for improving natural language

processing tasks in the Italian legal domain. Computer Law & Security Review, v. 52, 105908, 2024. https://doi.org/10.1016/j.clsr.2023.105908.

[12] LÖFFLER, C.; FREILE, A. M.; PIZARRO, T. R. Predicting potentially abusive clauses in Chilean terms of services with natural language processing. Artificial Intelligence and Law, 2025. https://doi.org/10.1007/s10506-025-09462-w.

[13] NIZZOLI, L.; AVVENUTI, M.; TESCONI, M.; CRESCI, S. Geo-semantic-parsing: AI-powered geoparsing by traversing semantic knowledge graphs. Decision Support Systems, v. 136, 113346, 2020. https://doi.org/10.1016/j.dss.2020.113346.

[14] PURVES, R. S.; CLOUGH, P.; JONES, C. B.; HALL, M. H.; MURDOCK, V. Geographic information retrieval: Progress and challenges in spatial search of text. Foundations and Trends in Information Retrieval, v. 12, n. 2-3, p. 164–318, 2018. http://dx.doi.org/10.1561/1500000034.

[15] RAMOS, G.; BATISTA, F.; RIBEIRO, R.; FIALHO, P.; MORO, S.; FONSECA, A.; GUERRA, R.; CARVALHO, P.; MARQUES, C.; SILVA, C. Leveraging transfer learning for hate speech detection in Portuguese social media posts. IEEE Access, v. 12, p. 101374-101386, 2024. https://doi.org/10.1109/ACCESS.2024.3430848.

[16] SHINGLETON, J.; BASIRI, A. Enhancing toponym identification: Leveraging Topo-BERT and open-source data to differentiate between toponyms and extract spatial relationships. AGILE: GIScience Series, v. 5, 12, 2024. https://doi.org/10.5194/agile-giss-5-12-2024.

[17] SILVA, M. J.; MARTINS, B.; CHAVES, M.; AFONSO, A. P.; CARDOSO, N. Adding geographic scopes to web resources. Computers, Environment and Urban Systems, v. 30, n. 4, p. 378–399, 2006. https://doi.org/10.1016/j.compenvurbsys.2005.08.003.

[18] SOUZA, C. N.; MARTÍNEZ-ARRIBAS, J.; CORREIA, R. A.; ALMEIDA, J. A.G.R.; LADLE, R.; VAZ, A. S.; MALHADO, A. C. Using social media and machine learning to understand sentiments towards Brazilian National Parks. Biological Conservation, v. 293, 110557, 2024. https://doi.org/10.1016/j.biocon.2024.110557.

[19] SOUZA, F. C.; NOGUEIRA, R. F.; LOTUFO, R. A. BERT models for Brazilian Portuguese: Pretraining, evaluation and tokenization analysis. Applied Soft Computing, v. 149, 110901, 2023. https://doi.org/10.1016/j.asoc.2023.110901.

[20] WU, Q.; YAO, P..; ZHU, H.; ZHU, W.; WU, Y.; LI, L. A deep learning approach to recognizing fine-grained expressway location reference from unstructured texts in Chinese. International Journal of Geographical Information Science, v. 38, n. 4, p. 654–674, 2024. https://doi.org/10.1080/13658816.2023.2301316.

[21] ZHOU, Z. Open-environment machine learning. National Science Review, v. 9, nwac123, 2022. https://doi.org/10.1093/nsr/nwac123.