PetroGeoNER: A Refined and Unified Dataset for NER in the Oil & Gas Domain

Higor Moreira[1], **Patricia Ferreira da Silva**[2], **Renata Vieira**[3], **Viviane Moreira**[1]

[1]Institute of Informatics - Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre – Brazil

{hmoreira, viviane}@inf.ufrgs.br

[2]Petrobras Research and Development Center (CENPES)
Rio de Janeiro - Brazil

patricia.fs@petrobras.com.br

[3]CIDEHUS - University of Évora
Évora - Portugal

renatav@uevora.pt

Abstract. *Named Entity Recognition (NER) is a task of Natural Language Processing (NLP) that deals with finding and categorizing relevant entities (*i.e.,* word n-grams) in a text, assigning them to predefined semantic categories. The availability of annotated datasets is crucial for developing NER models and assessing their quality. This becomes an issue considering underrepresented languages and specific domains. Furthermore, the word-level annotation required by NER datasets is laborious and prone to inconsistencies. Aiming to contribute to more resources for Portuguese, this paper compiled PetroGeoNER, a NER dataset in the Oil & Gas domain. The process of creating our dataset involved unifying, revising, and solving inconsistencies in two existing datasets. PetroGeoNER was used to train accurate NER models. Both the models and the dataset were made publicly available.*

1. Introduction

Named Entity Recognition (NER) is a task of Natural Language Processing (NLP) that deals with finding and categorizing relevant entities within a sentence (*i.e.,* word n-grams), grouping them into predefined categories such as person, organization, or location [Jurafsky and Martin, 2025]. NER is crucial in several downstream applications, such as information retrieval, question answering, text summarization, and knowledge graph construction [Keraghel et al., 2024].

The growing interest in the Oil & Gas industry, particularly regarding Brazil's pre-salt exploratory frontier, motivated the development of linguistic resources in Portuguese for this domain [Consoli et al., 2020, Gomes et al., 2021, Rodrigues et al., 2022, Cordeiro, 2024]. Specifically for NER, two datasets have been created: GeoCorpus [Amaral, 2017], a manually annotated dataset based on academic theses; and PetroNER [Freitas et al., 2023], a semi-automatically annotated dataset using technical bulletins from the Oil & Gas industry. Despite their contributions, these datasets still present some issues that may limit their applicability, such as annotation inconsistencies, variations in entity class definitions, and duplicate instances.

To address this gap and generate a larger, unified, and more consistent resource, we propose PetroGeoNER, a NER dataset for the Oil & Gas domain in Portuguese. PetroGeoNER combines the strengths of GeoCorpus and PetroNER, aligning their entity classes and standardizing the annotations using a multi-step process including entity mapping, correspondence identification, dataset refinement, and expert revision.

The main contributions of this paper include: (i) PetroGeoNER, a unified and refined NER dataset available at https://huggingface.co/datasets/hmoreira/PetroGeoNER, (ii) a methodology for aligning and consolidating NER datasets, including manual and automatic steps; and (iii) a trained model for NER of PetroGeoNER using a Transformer-based model, available at https://huggingface.co/hmoreira/xlm-roberta-large-petrogeoner.

2. Background

NER is a sequence labeling task in which the words (or tokens) from the input text are classified into predefined classes. The classes vary according to the domain of interest, but the methods for addressing the task tend to be the same.

In the Oil & Gas domain, NER is fundamental for extracting domain-specific entities from documents that contain information about geological formations, exploration sites, well operations, and historical production data, enabling the creation of knowledge graphs and information retrieval systems [Ittoo et al., 2016]. Typical entity categories in this context include rock types, sedimentary basins, geological periods, well identifiers, and fields, among others [Cordeiro et al., 2024].

The annotation of named entities follows the **BIO Tagging** scheme, which is one of the most commonly used approaches for sequence labeling in span-recognition problems. In this scheme, the tags capture both the boundary and the named entity type [Jurafsky and Martin, 2025]. In BIO tagging, the token that begins a named entity is labeled with 'B', the tokens that occur inside the named entity are tagged with an 'I', and any tokens outside the named entity are labeled with 'O'.

NER Methods cover a broad range of techniques from knowledge-based systems to deep learning architectures. Knowledge-based methods have been the foundational part of NER, originating from linguistic rules and lexical resources to identify named entities [Hanisch et al., 2005, Eftimov et al., 2017]. Conditional Random Fields (CRF) [Lafferty et al., 2001] have been widely applied for NER [McCallum and Li, 2003, Amaral, 2017]. They are probabilistic models that can capture the contextual dependencies between adjacent tokens, enabling more accurate prediction of entity labels by considering the relationship between the neighboring labels. More recently, deep learning methods that use neural networks to automatically learn representations of entities from the dataset have also been applied for NER [Huang et al., 2015, Gui et al., 2019]. Bidirectional Long Short-Term Memory (BiLSTM) networks process the labeling sequence in both directions, left-to-right and right-to-left, allowing them to capture the full sequence context around each token. BiLSTM can be combined with pre-trained word embeddings to enrich word representation. The introduction of Transformer-based models, such as BERT [Devlin et al., 2019], has improved the NER task by allowing the modeling of complex contextual relationships within text. Transformers use a self-attention mechanism to effectively capture contextual information. These encoders can be trained on large corpora, followed by fine-tuning on domain-specific NER datasets [Chalkidis et al., 2020, Lee et al., 2020] Although Large Language Models (LLMs) are showing improvements in several NLP tasks, the application of LLMs in NER has revealed some limitations, since LLMs are originally designed for text generation with some degree of liberty [Wang et al., 2023, Ashok and Lipton, 2023, Hu et al., 2023]. So far, encoder-based models are leading the field in the NER task. Empirical studies indicate that employing BERT as a classifier consistently outperforms traditional BiLSTM-CRF and the new LLMs [Keraghel et al., 2024].

Deep learning and Transformer-based architectures require large amounts of high-quality and consistent training data to train an accurate NER model. In this context, the creation and refinement of NER datasets are essential for NLP specialized domain applications.

3. Related Work

This section reviews existing NER datasets in the Oil & Gas domain and geological domains, focusing on dataset construction and refinement. We also briefly mention the types of models used to validate these datasets.

GeoCorpus [Amaral, 2017] is a NER dataset in Portuguese for the geological domain. It is based on theses and dissertations about Brazilian sedimentary basins. The dataset was manually annotated and revised by domain specialists. In their experiments, the authors used CRF, reporting an F1 score of 54%. GeoCorpus-2 [Consoli et al., 2020] is a revised version of the original dataset, including annotation correction, dataset conversion format, and a standardized split into training, validation, and test. They ran experiments using BiLSTM-CRF models and domain-enhanced embeddings, achieving an F1 score of 84.63%. GeoCorpus-3[Gomes et al., 2021] is a further refinement that removed duplicates, corrected annotations, and expanded the number of classes from 20 to 30 by splitting the class *outros* (others) into news classes. They reported an F1 score of 86% using Word2Vec and FastText models, only with the ten classes with the largest variety and highest number of named entities. More recently, Nunes et al. [2024a] identified and removed an additional duplicate, one ambiguous annotation, and corrected a typo in the class *sedimentaresSiliciclasticas* during their experiments.

PetroNER [Freitas et al., 2023, Cordeiro et al., 2024] is a NER dataset focused on the Oil & Gas domain in Portuguese, derived from 11 Petrobras' technical bulletins. The corpus was automatically annotated using a lexicon containing terms in the domain and rules defined by linguists. Following the automatic annotation, each instance was revised by domain specialists. They reported an F1 score of 82.8% using a fine-tuned BERTimbau [Souza et al., 2020] model.

In English, there is the OzRock dataset[Enkhsaikhan et al., 2021] built from reports from geological exploration of mineral resources in Western Australia. The dataset was automatically annotated using dictionary matching and an ensemble of four models. The test set was manually revised by domain experts. They report a F1 score of 82.19% using a BiLSTM model.

While important efforts have been made to develop NER datasets in the Oil & Gas and geological domain in Portuguese, they were developed independently, using distinct data sources and annotation guidelines. As a result, they differ in linguistic style, entity class definitions, and annotation granularity. Bridging these differences to create a unified and consistent resource remains an open challenge, which this work aims to address.

4. Dataset Creation

This work compiled PetroGeoNER, a refined and enlarged dataset for the Oil & Gas industry. PetroGeoNER combines the Technical Bulletins from PetroNER with theses and dissertations from GeoCorpus. Besides the need for identifying compatible entities in the source datasets, PetroGeoNER also includes a series of dataset refinement steps carried out by a domain specialist that made the dataset more consistent.

Figure 1 shows the pipeline used to create PetroGeoNER. Our inputs in Step 1 are the source datasets PetroNER and GeoCorpus. In Step 2, a domain specialist created a mapping scheme between the compatible entity classes in the source datasets. In Step 3, to identify correspondences, a compatibility list was created using the mapped entities and text edit distance. In Step 4, dataset refinement is performed to remove inconsistencies, and the dataset is split into training, validation, and test. In Step 5, the domain specialist thoroughly revised the test set to ensure the accuracy of the gold standard labels. The next sections describe in greater detail the source datasets and the process followed to create PetroGeoNER.

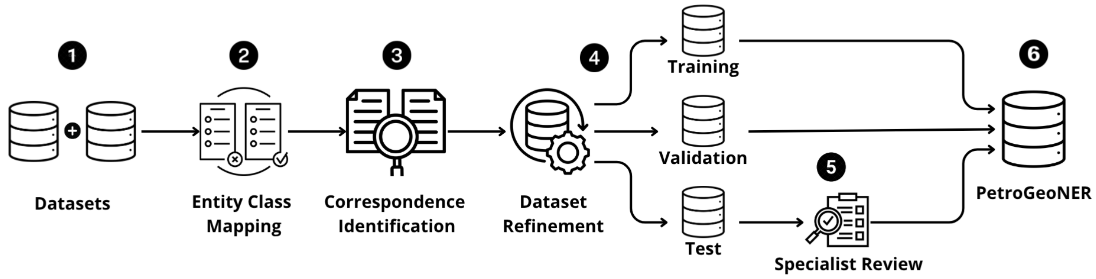

Figure 1. Pipeline for generating the PetroGeoNER dataset

4.1. Source Datasets

The two source datasets used to create PetroGeoNER were GeoCorpus-3 and PetroNER, which were already introduced in Section 2. PetroGeoNER contains 17,528 instances and 496,305 words, 24,902 named entities distributed along 13 classes.

GeoCorpus-3[1] is a revised version of GeoCorpus [Amaral, 2017], with more named entities, entity classes, instance corrections, and the split of the class '*outro*' (other) into new classes. GeoCorpus-3 consists of 19 texts, totaling 6,116 instances and 163,790 words, 8,954 named entities distributed along 30 classes Gomes et al. [2021].

[1] https://github.com/bsconsoli/GeoCorpus-V3

PetroNER[2] is a dataset in the Oil & Gas domain. It consists of 11 texts, totaling 24,035 instances, 615,418 words, and 18,757 named entities belonging to 18 classes Cordeiro et al. [2024].

4.2. Entity Class Mapping

The first step in combining the datasets was to map their classes. We started by compiling all the descriptions and definitions from the source datasets to create a scheme with compatible entity classes across the datasets. In this process, we identified the entity classes from GeoCorpus-3 as finer-grained entities (with more specific subclasses) and the classes in PetroNER are coarser-grained. Using this scheme, we mapped multiple entity classes across the datasets. For example, the entity classes from GeoCorpus-3 for geological chronological units (*ano, periodo, epoca, era,* and *eon*) were mapped into the entity class *UNIDADE_CRONO* (chronological unit) from PetroNER. The entire entity class mapping process was revised manually by a domain specialist.

Figure 2 illustrates the mapping of entity classes between PetroNER and GeoCorpus. `PetroGeoNER`, inherited the entity classes from PetroNER. In orange, we represent the original entities from PetroNER; in green, the GeoCorpus entities that were reassigned to different classes; and in blue, new classes created for `PetroGeoNER`, which are included in GeoCorpus and were standardized. The classes in blue were created based on suggestions from the specialist domain, as those entities are also present in PetroNER but were not previously annotated.

4.3. Correspondence Identification

The correspondence identification evaluates the compatibility between the entity instances in GeoCorpus-3 and PetroNER. Using the entity class mapping, we compiled a list containing the unique occurrences of each entity from the source datasets. The entities in the original datasets were matched using the list of entities. Figure 2 illustrates which entity classes were used in the correspondence identification step.

The authors of PetroNER compiled a lexicon for each class containing many possible named entities for that class. These lexica were used to annotate PetroNER, and we used them to identify correspondences between PetroNER and GeoCorpus-3. To identify correspondences, each GeoCorpus-3 entity is compared with all the instances from the lexicon. The pair with the highest similarity (according to the edit distance) was considered a match if the score was above 0.85. If the pair with the highest similarity had a score below the threshold, the domain specialist checked it to decide whether it was a match.

4.4. Dataset Refinement

During the Entity Class Mapping (Section 4.2) and Correspondence Identification (Section 4.3), we identified inconsistencies in the instances from the source datasets. The dataset refinement process was performed to remove these inconsistencies. Next, we enumerate the problems and the solutions we adopted. Table 1 presents the results from the filtering of the source datasets.

1. **Duplicate instances**, *i.e.,* identical sentences, can lead to data leakage and should be discarded from the datasets Nunes et al. [2024b]. Despite being in its third

[2] `https://petroles.ica.ele.puc-rio.br/`

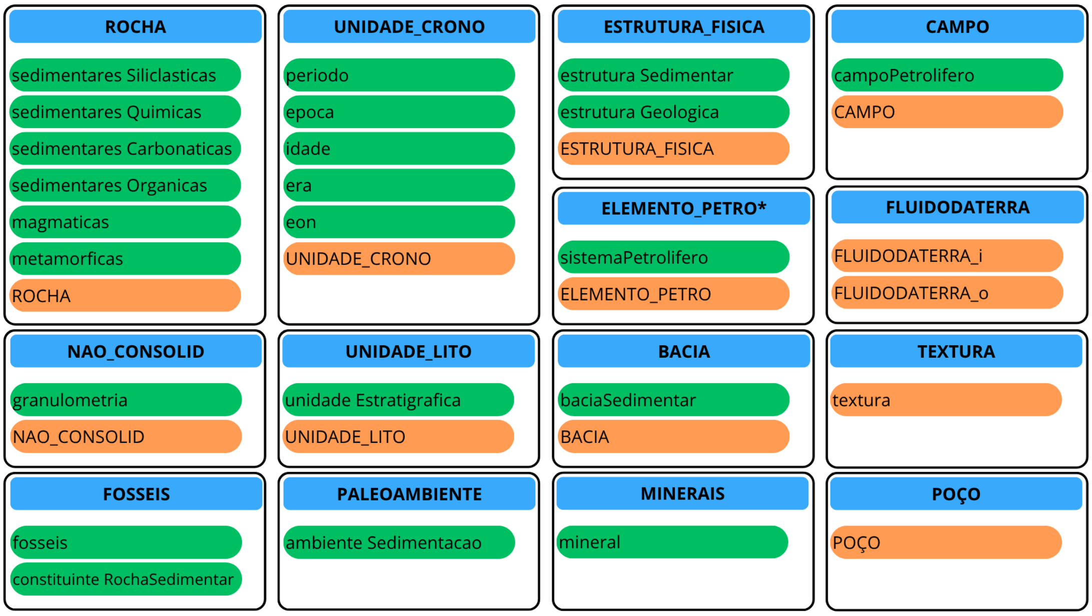

Figure 2. Mapping of entity classes from GeoCorpus-3 and PetroNER. The `PetroGeoNER` entity classes resulting from the mapping process are shown in blue. Green indicates entities originally from GeoCorpus-3, and orange indicates those from PetroNER. The class `ELEMENTO_PETRO` was removed during the dataset refinement process.

version, GeoCorpus-3 still contains duplicate instances, as reported by Nunes et al. [2024a], who checked for exact matches. In the PetroNER and GeoCorpus-3 dataset, to identify duplicates, we applied the edit distance metric, with a threshold of 0.98. Instances with exact matches were automatically removed, and the instances above the threshold were manually revised.

2. **Non-informative and noisy instances**, mostly originating from footnotes, headers, and references. PetroNER authors did not delimit which part of the original document should be included to compose the dataset instances. Additionally, some documents were digitized using OCR, which may have introduced extraction errors. To identify and remove these instances, we defined a set of rules and regular expressions. Instances matching all three criteria were removed from the resulting dataset: (i) instances without any annotated named entities, (ii) instances with fewer than 25 tokens, and (iii) instances in which more than 30% of the characters are numbers, punctuation, special characters, or long sequences of uppercase letters. We restricted the removal to instances without any named entities to avoid reducing the number of named entities annotated.
3. **Entity mislabeling** refers to the incorrect assignment of a named entity. This includes assigning the wrong entity class or tagging the boundaries incorrectly. The most common cases of incorrect sequence labeling were identified as *over-specification*. This refers to the use of detailed or context-specific descriptions in a labeled entity that go beyond the general classification. For example, within the entity class *ROCHA*, the entity '*Arenito Marlim*' is considered an over-specification, because it includes specific terms that are not necessary for the entity

class. In this case, the correction applied is labeling the sequence as '*Arenito*' only, without '*Marlim*'.

4. **Generic annotation** are the opposite of over-specification. Generic entities are words commonly found in the sequence labeling of a named entity, but when isolated, they do not carry meaning to be classified[3]. Examples of generic annotations are given in Table 2.

Table 1. Number of instances in the source datasets before and after refinement

	instances		
Dataset	**#Original**	**#Filtered**	**Δ**
GeoCorpus-3	6,116	6,028	-88
PetroNER	24,035	14,414	-9,621

Table 2. Examples of generic tokens misclassified as a named entity. The generic tokens that do not refer to a specific named entity are underlined.

Entity Class	**Token**	**Instance**
BACIA	bacia	Todavia, nem todos os 430 poços escolhidos atingiram o embasamento da bacia.
ROCHA	rocha	Esta rocha possui logfácies com baixa radioatividade, altas resistividade e densidade.
POÇO	poço	Este poço atingiu uma profundidade total de 1096m, atravessando 12 níveis vulcânicos.
NAO_CONSOLID	sedimentos	Os sedimentos foram depositados entre o Neotriássico e o final do Cretáceo
CAMPO	campo	Neste campo, o reservatório foi divido em três unidade sismoestratigráficas de alta resolução.
ESTRUTURA_ FISICA	estrutura	Nessa estrutura são reconhecidos elementos de compressão, tais como falhas reversas e dobras.

The dataset refinement process also removed classes from the source datasets. In GeoCorpus-3, we removed the following classes: *contextoGeologicoDeBacia, estratigrafia, procedimentoMetodologico, planctonico, geomorfologia, unidadeGeotectonica, bentonico,* and *elementoQuimico*. These classes were introduced in GeoCorpus-3 and do not have the criteria and description used for their categorization, different from the 20 original classes in [Amaral, 2017].

In PetroNER, we removed *POÇO_Q, POÇO_T, POÇO_R, TIPO_POROSIDADE, EVENTO_PETRO, ELEMENTO_PETRO,* and *FLUIDO*. These classes had few occurrences, or the annotation was problematic. The *POÇO* specification, *TIPO_POROSIDADE*, and *FLUIDO* have very few named entities after the refinement. *EVENTO_PETRO* and *ELEMENTO_PETRO* were removed due to annotation problems. According to our domain specialist, the tokens do not cover the entire scope, and the classes share similar tokens, making it hard to revise and annotate correctly. In addition, *FLUIDODATERRA_i* and *FLUIDODATERRA_o* have a similar problem of sharing tokens, but we opted to merge them into a unique class *FLUIDODATERRA*.

[3] Further NLP processing, such as coreference resolution, could be applied in order to add context to these generic expressions.

Table 3 presents the results of the dataset refinement process and summarizes the statistics of `PetroGeoNER`. The dataset was divided into three splits: 70% for training, 10% for validation, and 20% for testing. To ensure a consistent class distribution, we preserved the original proportion of entity classes across the splits, allowing a maximum variation of 5 percentage points per class relative to the full dataset. Additionally, we balanced the splits by data source, ensuring that instances from both GeoCorpus-3 and PetroNER were proportionally represented in all partitions.

`PetroGeoNER` does not include the instances from GeoCorpus-3 that do not contain any named entities, for two main reasons: (i) to avoid introducing sentences that, although revised, were not specifically reviewed to confirm whether they contain entities from PetroNER classes; and (ii) PetroNER already contains a sufficient number of non-entity instances, which were revised for the three new classes introduced in `PetroGeoNER`: *FLUIDODATERRA*, *FOSSEIS*, and *PALEOAMBIENTE*. The number of removed sentences in PetroNER is higher than in GeoCorpus-3 because the latter is in its third version and has undergone three rounds of revisions. PetroNER, on the other hand, had not undergone any revisions until now.

Table 3. Results from the dataset refinement process in the source datasets. Original and Filtered refer to the original number of named entities and the updated value after the refinement, respectively. Δ denotes the difference between the original and filtered values. Additionally, the table includes the number of named entities from `PetroGeoNER`.

	GeoCorpus-3			**PetroNER**			**PetroGeoNER**
Entity Class	**#Original**	**#Filtered**	Δ	**#Original**	**#Filtered**	Δ	**Named Entities**
BACIA	552	304	-248	4,057	2,585	-1,472	2,889
CAMPO	6	6	0	708	495	-213	501
ESTRUTURA_FISICA	164	198	34	2,042	1,702	-340	1,900
FLUIDODATERRA	0	0	0	1620	1,621	1	1,621
FOSSEIS	132	252	120	0	1,244	1,244	1,496
MINERAIS	212	242	30	0	647	647	889
NAO_CONSOLID	129	195	66	1,036	360	-676	555
PALEOAMBIENTE	146	221	75	0	1,900	1,900	2,121
POÇO	0	0	0	1,230	385	-845	385
ROCHA	2,451	2,112	-339	2,774	2,326	-448	4,438
TEXTURA	0	0	0	141	138	-3	138
UNIDADE_CRONO	2,870	2,885	15	2,920	2,868	-52	5,753
UNIDADE_LITO	764	727	-37	1,492	1,489	-3	2,216
Total	7,426	7,142	-284	18,020	17,883	-137	24,902

4.5. Specialist Review

To create a gold standard, the domain specialist manually revised the test split, following a procedure similar to Enkhsaikhan et al. [2021]. This manual review served two purposes: (i) to assess and validate the quality of the dataset creation and refinement process; and (ii) to produce a test set with high annotation reliability for model evaluation. During the review, the specialist made corrections, adding and removing annotations as needed.

In this step, we employed the text-annotation environment INCEpTION[4]. The environment provides a recommendation system to help the annotator, suggesting words that have already been annotated as a named entity.

[4] `https://inception-project.github.io/`

5. Training and Evaluating NER models

In addition to compiling the dataset, we also trained NER models, aiming to make them publicly available. This section describes the main results of the NER experiments in the PetroGeoNER dataset. For training the NER models, we used five models based on BERT [Devlin et al., 2019, Souza et al., 2020] and RoBERTa [Conneau et al., 2019]. The variations include multilingual and Portuguese models. We used the following hyperparameters for training: evaluation_strategy = epochs, learning_rate = 3e-5, num_train_epochs = 10, weight_decay = 0.01, train and eval batch_size = 16, optimizer = Adam. The remaining parameters follow the default suggested by HuggingFace[5].

Table 4. Evaluation results of NER models trained on the PetroGeoNER dataset. The best results in each metric are in bold.

Models	Parameters	Language	F1-Micro	F1-Macro
BERT Base	110M	Multilingual	0.89	0.86
BERTimbau Base	110M	Portuguese	0.88	0.85
BERTimbau Large	335M	Portuguese	0.89	**0.87**
XLM-RoBERTa Base	125M	Multilingual	0.89	0.86
XLM-RoBERTa Large	561M	Multilingual	**0.90**	**0.87**

To evaluate the models, we computed the following metrics: precision, recall, F1-score micro and macro, using the SeqEval library[6]. The library is supported by HuggingFace and calculates the results based on the sequence of tags using the BIO scheme. The experiments were done on a computer with 224 CPU nodes equipped with 2 Intel Xeon 6248 processors and 384GB of RAM, as well as 24 GPU nodes that each have 8 Nvidia Tesla V100 32GB with 754GB of RAM. We chose the Python 3.10 programming language because of its variety of libraries for NLP.

Table 4 shows the models used in the experiments, the number of parameters, the main language, and the scores of F1, both micro and macro-averaged. The results indicate that the model with the best performance is **XLM-RoBERTa Large**, achieving an F1 Micro score of 0.9, and tied with BERTimbau Large in the F1 Macro metric, both scoring 0.87. In comparison to NER results reported in related work (even though they are not directly comparable, due to the different number of instances, entities, and classes), our models yielded better scores. Our F1-micro of 0.90 is higher than the 0.89 achieved by Nunes et al. [2024a] on GeoCorpus-3 using XLM-RoBERTa-Large with CRF and cross-validation, and higher than the score of 0.83 reported by Cordeiro [2024] on PetroNER using BERTimbau Large.

Table 5 shows the results for XLM-RoBERTa Large (*i.e.,* the best model) for each entity class. We can see F1-scores ranging from 0.77 to 0.95. The classes with the lowest F1-scores tend to have fewer instances or high variability, which makes it harder for the model to generalize during prediction. In many classes, the scores for recall and precision are similar. However, recall is lower than precision in *NAO_CONSOLID* and *PALEOAMBIENTE*. This means that the model failed to recognize entities belonging to those classes.

[5] https://huggingface.co/docs/transformers/en/tasks/token_classification

[6] https://github.com/chakki-works/seqeval

Table 5. The results from the XLM-RoBERTa Large model in the test split. The table shows the precision, recall, F1-score, and support for each entity class.

Entity Class	Precision	Recall	F1-score	Support
BACIA	0.91	0.96	0.94	581
CAMPO	0.87	0.81	0.84	99
ESTRUTURA_FISICA	0.89	0.84	0.86	396
FLUIDODATERRA	0.89	0.85	0.87	339
FOSSEIS	0.90	0.76	0.82	336
MINERAIS	0.93	0.83	0.88	217
NAO_CONSOLID	0.89	0.69	0.78	131
PALEOAMBIENTE	0.85	0.71	0.77	486
POÇO	0.97	0.92	0.94	84
ROCHA	0.93	0.93	0.93	848
TEXTURA	0.88	0.79	0.84	29
UNIDADE_CRONO	0.95	0.96	0.95	1119
UNIDADE_LITO	0.91	0.88	0.90	468

6. Conclusion

In this work, we introduced `PetroGeoNER`, a revised and unified dataset for NER in the Oil & Gas domain in Portuguese. The creation integrated two datasets: GeoCorpus-3 and PetroNER. The dataset creation process included entity class mapping, correspondence identification, refinement, and a gold standard test split in which all instances have been manually revised by a domain specialist. Beyond dataset creation, we also fine-tuned and evaluated several state-of-the-art NER language models, with XLM-RoBERTa Large achieving the best performance, being closely followed by BERTimbau Large.

`PetroGeoNER` represents a step forward for NLP applications in the Oil & Gas industry. It provides a reliable benchmark for future research and resources to support domain-specific tasks such as information extraction and building knowledge graphs.

Despite its contributions, some limitations remain. The manual labeling of an NER dataset is very laborious. Thus, currently, only the test split was manually revised by a domain specialist. As a result, there may be inconsistencies in the training and validation splits. Additionally, some classes have few training examples, potentially impacting model performance in those categories.

In future work, we plan to improve dataset quality by extending the manual revision process to additional data splits and performing error analysis on entity classes with F1-scores below 0.8. This will help identify error patterns for better annotation adjustments and model performance. We also aim to implement data augmentation for these categories and evaluate how large language models compare to models such as BERT and RoBERTa.

7. Acknowledgments

This work has been partially fundend by CENPES Petrobras, CNPq-Brazil, and Capes Finance Code 001.

References

Daniela Oliveira Ferreira Amaral. *Reconhecimento de entidades nomeadas na Área da geologia: bacias sedimentares brasileiras*. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 2017. URL http://tede2.pucrs.br/tede2/handle/tede/8035.

Dhananjay Ashok and Zachary C Lipton. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*, 2023.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

Bernardo Consoli, Joaquim Santos, Diogo Gomes, Fabio Cordeiro, Renata Vieira, and Viviane Moreira. Embeddings for named entity recognition in geoscience portuguese literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4625–4630, 2020.

Fábio Corrêa Cordeiro. *Petro KGraph: a methodology for extracting knowledge graph from technical documents - an application in the oil and gas industry*. PhD thesis, Fundação Getulio Vargas, Escola de Matemática Aplicada, 2024. URL https://hdl.handle.net/10438/35868.

Fábio Corrêa Cordeiro, Patrícia Ferreira da Silva, Alexandre Tessarollo, Cláudia Freitas, Elvis de Souza, Diogo da Silva Magalhaes Gomes, Renato Rocha Souza, and Flávio Codeço Coelho. Petro nlp: Resources for natural language processing and information extraction for the oil and gas industry. *Computers & Geosciences*, 193: 105714, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, June 2019.

Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488, 2017.

Majigsuren Enkhsaikhan, Wei Liu, Eun-Jung Holden, and Paul Duuring. Auto-labelling entities in low-resource text: a geological case study. *Knowl. Inf. Syst.*, 63(3):695–715, March 2021. ISSN 0219-1377. doi: 10.1007/s10115-020-01532-6. URL https://doi.org/10.1007/s10115-020-01532-6.

Cláudia Freitas, Elvis Souza, Maria Clara Castro, Tatiana Cavalcanti, Patricia Ferreira da Silva, and Fábio Corrêa Cordeiro. Recursos linguísticos para o pln específico de domínio: o petrolês. *Linguamática*, 15(2):51–68, Dez. 2023. doi: 10.21814/lm.15.2.412. URL https://linguamatica.com/index.php/linguamatica/article/view/412.

Diogo da Silva Magalhães Gomes, Fábio Corrêa Cordeiro, Bernardo Scapini Consoli, Nikolas Lacerda Santos, Viviane Pereira Moreira, Renata Vieira, Silvia Moraes, and

Alexandre Gonçalves Evsukoff. Portuguese word embeddings for the oil and gas industry: Development and evaluation. *Computers in Industry*, 124:103347, 2021.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. Cnn-based chinese ner with lexicon rethinking. In *ijcai*, volume 2019, pages 4982–4988, 2019.

Daniel Hanisch, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck. Prominer: rule-based protein and gene entity recognition. *BMC bioinformatics*, 6:1–9, 2005.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*, 2023.

Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Ashwin Ittoo, Antal van den Bosch, et al. Text analytics in industry: Challenges, desiderata and trends. *Computers in Industry*, 78:96–107, 2016.

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

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. Recent Advances in Named Entity Recognition: A Comprehensive Survey and Comparative Study. 2024.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA, 2001.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 188–191, 2003.

Rafael O. Nunes, Andre S. Spritzer, Dennis G. Balreira, Carla M. D. S. Freitas, and Joel L. Carbonera. An evaluation of large language models for geological named entity recognition. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 494–501, 2024a. doi: 10.1109/ICTAI62512.2024.00076.

Rafael Oleques Nunes, André Susliz Spritzer, Carla Maria Dal Sasso Freitas, and Dennis Giovani Balreira. Reconhecimento de entidades nomeadas e vazamento de dados em textos legislativos. *Linguamática*, 16(2):141–166, 2024b.

Rafael BM Rodrigues, Pedro IM Privatto, Gustavo José de Sousa, Rafael P Murari, Luis CS Afonso, João P Papa, Daniel CG Pedronette, Ivan R Guilherme, Stephan R Perrout, and Aliel F Riente. Petrobert: a domain adaptation language model for oil and gas applications in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 101–109. Springer, 2022.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*, 2023.