




# Sequence Labeling in Product Descriptions on Invoices: Comparing LLM-based settings with a CRF baseline

Eduardo Darrazão   [ Universidade Tecnológica Federal do Paraná | [darrazao@alunos.utfpr.edu.br](mailto:darrazao@alunos.utfpr.edu.br) ]

Krerley Oliveira  [ Universidade Federal de Alagoas | [krerley@mat.ufal.br](mailto:krerley@mat.ufal.br) ]

Luiz Celso Gomes-Jr   [ Universidade Tecnológica Federal do Paraná | [lcjunior@utfpr.edu.br](mailto:lcjunior@utfpr.edu.br) ]

 DAINF, Universidade Tecnológica Federal do Paraná, Av. Sete de Setembro, 3165, Rebouças, Curitiba, PR, 80230-901, Brazil.

**Received:** 24 March 2025 • **Accepted:** 25 July 2025 • **Published:** 24 October 2025

**Abstract** Electronic invoices are present in most commercial transactions since several countries require their issue in the purchase, sale, and transportation of goods. The accurate identification of elements within these invoices is crucial for governmental oversight, aiding in tasks such as detecting overpricing in public contracts. However, this identification is a challenge due to the diversity of products, as well as variations and errors in filling out the information. This article aims to compare the performance of a model developed using a traditional Conditional Random Fields (CRF) technique for the task with models based on large language models adapted for this task. The goal is to assess whether language models can be effectively used to improve the performance in this scenario. The paper assesses the use of several modeling approaches, including the influence of language in the base model (Portuguese-specific vs. Multilingual BERT), as well as alternatives for the classification head (fine-tuning with a linear layer vs. feature-extraction with BiLSTM and a linear layer, with or without a CRF layer). The best model, which combines a Portuguese BERT-based approach with a Conditional Random Fields layer, achieves an F1-score improvement of approximately 4% over the baseline model that relies solely on CRF. The tests used data from invoices issued in Brazil in 2021 in the context of public contracts.

**Keywords:** Sequence Labeling, Named Entity Recognition, Invoices, Large Language Models, BERT, Conditional Random Fields, Brazilian Portuguese

© Published under the Creative Commons Attribution 4.0 International Public License (CC BY 4.0)

## 1 Introduction

Every day millions of companies and individuals throughout Brazil carry out operations involving the purchase, sale, and transportation of goods [F. N. de Oliveira, 2020]. These operations are required to issue electronic invoices, which are stored by the government for various purposes. One important use for such information is the identification of frauds. Tax fraud results in annual losses of approximately 400 billion Reals, according to the Brazilian Institute of Planning and Taxation, where ICMS (stands for Tax on Circulation of Goods and Services) represents a significant portion of them [F. N. de Oliveira, 2020].

To enable the automation of the analysis of invoices, it is necessary to adequately identify the traded products. However, a large amount of useful information for product identification is in the textual description field, which is freely written by the issuer of the invoice. In the case of electronic invoices related to medicines, the text in their descriptions contains hierarchical subclasses representing the composition and presentation of the medications, which may include the brand name or generic name of the medication, dosage, product quantity, and pharmaceutical form. These elements correspond to named entities such as Product, Presentation, Packaging, Quantity, Concentration, Content, and Additional

information (e.g., syringes, cups). These contents can be written with different terminologies for the same element, with abbreviations, with or without special characters. Furthermore, there may be incorrect entries or even items that are not medications. This variability makes sequence labeling, a task aimed at assigning a categorical label to each token in a sequence, essential for structuring this information.

The diversity and unpredictability within these textual descriptions demand an approach that can interpret and organize such data effectively. This necessity aligns with the domain of Natural Language Processing (NLP), which specializes in developing computational methods for processing and analyzing human language. More specifically, the NLP task of Sequence Labeling focuses on assigning labels to a category of morphemes that generally have similar grammatical properties, which is the basis for Named Entity Recognition and related tasks [He *et al.*, 2020; Jurafsky and Martin, 2008]. This is crucial for transforming unstructured product descriptions into structured data suitable for automated analysis.

Large Language Models (LLMs) based on Transformers have had a significant impact on Natural Language Processing. They are capable of capturing context and dependencies between words by learning from large amounts of text [Bender *et al.*, 2021]. This makes them well-suited for handling the varied language, abbreviations, and complexities found

in invoice descriptions.

LLMs like Generative Pre-trained Transformer (GPT) [Radford and Narasimhan, 2018] and Bidirectional Encoder Representations from Transformers (BERT) [Devlin *et al.*, 2019] are highly adaptable and can be fine-tuned or used in feature-extraction setups to focus specifically on identifying products and their characteristics within these invoices, potentially making them effective at this task, and possibly boosting the accuracy and efficiency of identifying products within these documents.

In this article, two main sequence labeling approaches are applied to the product descriptions: the first is based solely on a Conditional Random Fields (CRF) model, as developed in Darrazão *et al.* [2023], and the second involves adapted versions of two BERT language models (BERTimbau, a Portuguese version [Souza *et al.*, 2019, 2020a], and a Multilingual BERT) along with classifiers. The first model serves as a baseline to assess whether language models can be effectively used to improve the performance of this task.

The remainder of this paper is organized as follows: Section 2 briefly reviews key concepts relevant to this work; Section 3 discusses related work on fiscal documents analysis and Sequence Labeling; Section 4 details the methods and model architectures used throughout this article; Section 5 presents and discusses the results; and Section 6 provides the final remarks of this work.

## 2 Background Concepts

This section briefly reviews fundamental concepts essential for understanding the methods and models discussed in this paper.

### 2.1 Natural Language Processing and Sequence Labeling

Natural Language Processing (NLP) is an interdisciplinary field, at the intersection of computer science and linguistics, focused on enabling computers to process and understand human language [Jurafsky and Martin, 2008]. A core task within NLP is Information Extraction, which aims to automatically extract structured information from unstructured or semi-structured text [Seymore and Rosenfeld, 1999]. Sequence Labeling is a specific type of information extraction task where each token in a sequence (e.g., words in a sentence) is assigned a categorical label. This is fundamental for tasks like Named Entity Recognition (NER), which identifies and categorizes named entities (e.g., persons, organizations, locations, product names) in text [He *et al.*, 2020; Jurafsky and Martin, 2008].

Representing text numerically is crucial for machine learning models. Traditional methods include count-based vector representations like TF-IDF. More recent approaches involve *word embeddings*, which are dense vector representations that capture semantic relationships between words. Proximity between vectors often correlates with semantic similarity [Mikolov *et al.*, 2013; Pennington *et al.*, 2014; Bender *et al.*, 2021]. Contextualized word embeddings, generated by models like BERT, provide different representations for a token

depending on its context [Devlin *et al.*, 2019].

## 2.2 Key Models

### Conditional Random Fields (CRF)

CRFs are probabilistic graphical models used for segmenting and labeling sequential data [John Lafferty, 2001]. They model the conditional probability of a label sequence given an observation sequence. Unlike Hidden Markov Models (HMMs), CRFs can incorporate a wide range of arbitrary, overlapping features from the observation sequence and can model dependencies between labels directly. Linear-chain CRFs, commonly used in NLP, consider dependencies between adjacent labels and can work bidirectionally by considering the entire input sequence to make predictions for each token [Sutton *et al.*, 2012].

### Long Short-Term Memory (LSTM)

LSTMs are a type of recurrent neural network (RNN) effective at capturing long-range dependencies in sequential data [Hochreiter and Schmidhuber, 1997]. Bidirectional LSTMs (BiLSTMs) process sequences in both forward and backward directions, allowing them to consider past and future context at each time step, which is beneficial when the entire sequence is available, though not always applicable in real-time streaming scenarios where future context is unavailable.

### Transformers

The Transformer architecture, introduced by Vaswani *et al.* [2017], relies on self-attention mechanisms to weigh the importance of different words in a sequence when processing text. This allows for parallel processing and has led to significant advances in NLP, forming the basis for models like BERT and GPT.

### Large Language Models (LLMs)

LLMs, such as BERT [Devlin *et al.*, 2019] and GPT [Radford *et al.*, 2019], are pre-trained on vast amounts of text data. BERT (*Bidirectional Encoder Representations from Transformers*) is particularly relevant to this work. It uses the Transformer encoder architecture to learn contextual representations of words. BERT's utility, particularly the BERTimbau variant [Souza *et al.*, 2020a] pre-trained on Brazilian Portuguese text, stems from its ability to be fine-tuned for specific downstream tasks like sequence labeling, its manageable size for deployment compared to larger models, and its demonstrated effectiveness in Portuguese.

## 2.3 Evaluation Metrics for Sequence Labeling

The performance of sequence labeling models is typically evaluated using metrics calculated based on true positives (TP), false positives (FP), and false negatives (FN) at the entity or token level. For multi-class problems, these are often micro-averaged or macro-averaged. The primary metrics used are [Jurafsky and Martin, 2008]:

- **Precision:** The proportion of correctly identified positive instances among all instances identified as positive ( $TP / (TP + FP)$ ).
- **Recall:** The proportion of correctly identified positive instances among all actual positive instances ( $TP / (TP + FN)$ ).
- **F1-Score:** The harmonic mean of precision and recall ( $2 * (Precision * Recall) / (Precision + Recall)$ ), providing a single measure of performance.
- **Accuracy:** The proportion of correctly labeled tokens (both positive and negative) out of the total number of tokens. While common, it can be misleading for imbalanced datasets.

These metrics are standard in NER evaluation frameworks like CoNLL [Tjong Kim Sang and De Meulder, 2003] and tools like SEQeval.

## 3 Related Work

### 3.1 Applications in the Fiscal Domain

There are several works related to the analysis of fiscal documents to inspect and prevent fraud, which differ in both techniques and types of documents. In general, these works focus on the detection of banking fraud or business transactions, not necessarily using invoice data but rather general accounting data.

In the governmental context, the Brazilian Federal Court of Audit (TCU) stands out in developing projects using Data Science and Artificial Intelligence to collaborate in preventing and detecting fraud and corruption [Veras Carvalho Menezes, 2022], such as ALICE (Bidding, Contracts, and Notices Analyzer); MONICA (Integrated Monitoring for Procurement Control); SOFIA (Guidance System on Facts and Evidence for Auditors); ADELE (Analysis of Disputes in Electronic Bidding).

ALICE, among other systems, aims to assist relevant authorities in controlling public accounts. According to TCU (2017) [Veras Carvalho Menezes, 2022]:

*the ALICE system has enabled timely and automated evaluation of bidding notices and auction minutes, identifying signs of irregularities, fraud, embezzlement, and waste of public resources, enabling more efficient and effective control actions.*

The ALICE system is applied to a massive amount of data of different natures, including healthcare products. The report by TCU discusses the need to structure this data. The aim of this article is to assist with such demands, potentially improving the performance of systems like ALICE and expanding them to new areas.

With the purpose of examining products with overpricing and facilitating the control and prevention of undue public expenses, [Pereira, 2020] proposes heterogeneous networks for classifying products in electronic invoices for public procurement. This work presents a similar objective to this article; however, it employs different techniques: modeling the dataset in a bipartite network, and using a supervised inductive algorithm based on heterogeneous bipartite networks for

text classification. Works like this help us expand the scope of the problem and demonstrate other methods to try to solve it. It is conceivable that the combination of the strategies proposed here with those presented by the authors of Pereira [2020] may provide even better results.

### 3.2 Sequence Label Models and Other Applications

Considering other applications of Sequence Labeling, the Conference on Computational Natural Language Learning (CoNLL) is an annual conference that features shared tasks and research in machine learning for natural language processing. In 2003, evaluation methodologies for Named Entity Recognition were discussed [Tjong Kim Sang and De Meulder, 2003], using different techniques such as Maximum Entropy models and Hidden Markov Models. Although it is not exactly the same task, similarities can be found in the methods employed, such as the use of archetypes for data annotation and the separation of features into groups, as done in the base CRF model.

The Yet Another Sequence Tagger (YASET) is a Sequence Labeling tool designed for biomedicine texts. [Tourille *et al.*, 2018] present the tool, which is evaluated in part-of-speech and Named Entity Recognition tasks. The algorithm is built in multiple layers, combining artificial neural networks with Conditional Random Fields (CRFs), and using different strategies for feature construction (such as word embedding, for example) and model optimization. We can observe that there are similarities in the strategies employed, such as the use of archetypes for annotations and the use of complex models with a CRF layer, as seen in Souza *et al.* [2019].

The evolution of sequence labeling has seen a decisive shift from traditional statistical models, which depend on extensive feature engineering, towards end-to-end neural architectures capable of automatic feature learning [Bose *et al.*, 2021]. While effective, models like Conditional Random Fields (CRFs) required significant domain-specific manual effort [Huang *et al.*, 2015]. This challenge is particularly pronounced in specialized domains such as clinical medicine, where the complexity of terminology, ambiguity, and the high cost of creating large-scale annotated datasets necessitate more advanced and adaptable approaches. A survey by Bose *et al.* [2021] charts this progression, highlighting the unique difficulties in clinical Named Entity Recognition (NER), such as handling nested entities and the critical role of context, which has driven the field towards deep learning solutions.

A key advancement was the hybridization of neural networks with traditional graphical models. The Bidirectional Long Short-Term Memory-CRF (BiLSTM-CRF) architecture became a dominant baseline, combining the rich, bidirectional contextual representations from LSTMs with the structural constraints imposed by a CRF layer to ensure valid label sequences [Huang *et al.*, 2015]. With the advent of Transformer networks, this paradigm evolved further. The work of Souza *et al.* [2020b] exemplifies this shift, demonstrating that a BERT-CRF model could establish new state-of-the-art performance for Portuguese NER. This approach validates a crucial insight: even with the powerful contextual embed-

dings from a pre-trained model like BERT, the CRF layer remains highly valuable for modeling label dependencies and enforcing global sequence-level optimality, effectively combining the strengths of both deep representation learning and structured prediction.

The performance of these models is further amplified through domain adaptation, a critical step for achieving high accuracy in specialized fields. General-purpose language models often struggle with the distinct vocabularies and linguistic patterns found in areas like biomedicine. The study by Ganiz *et al.* [2022] provides a quantitative analysis of this effect. They show that a BiLSTM-CRF model using embeddings from BioBERT—a BERT model further pre-trained on a large biomedical corpus—achieves an F1-score improvement on biomedical NER tasks compared to the same architecture using generic BERT or static word embeddings. This underscores the efficacy of a multi-stage transfer learning strategy, where general pre-training is followed by domain-adaptive pre-training and subsequent task-specific fine-tuning, offering a practical pathway to high performance without requiring massive, labeled in-domain datasets from scratch.

Concurrently, for information extraction from semi-structured, visually-rich documents like invoices, the field is moving beyond pure sequence labeling towards generative and multimodal paradigms. This task is distinct from traditional NER as it must account for document layout and visual cues. An example of such approach is presented by Cao *et al.* [2023] with their GenKIE model. This framework reframes the task from an extractive one (classifying tokens) to a generative one (producing a structured output). GenKIE, a multimodal sequence-to-sequence model, leverages visual, layout, and textual features to directly generate key-value pairs. This generative approach offers advantages like inherent robustness to Optical Character Recognition (OCR) errors and a simplified annotation process that does not require laborious token-level tagging.

## 4 Methodology

The problem addressed is to classify parts of medical product descriptions in invoices into 7 distinct, non-intersecting categories. Furthermore, the descriptions may not necessarily contain all the categories distributed in their parts. The considered categories are:

1. Product: the product being sold.
2. Presentation: how the product is presented (tablet, capsule, solution, etc).
3. Packaging: how the product is packaged (box, ampule).
4. Quantity: the quantity of products sold.
5. Concentration: the concentration of the product (in milligrams, percentage, etc).
6. Content: related to the weight/quantity/volume of the product.
7. Additional: additions that come with the product (syringes and cups, for example).

Figure 1 shows statistics for the labels in the annotated dataset.

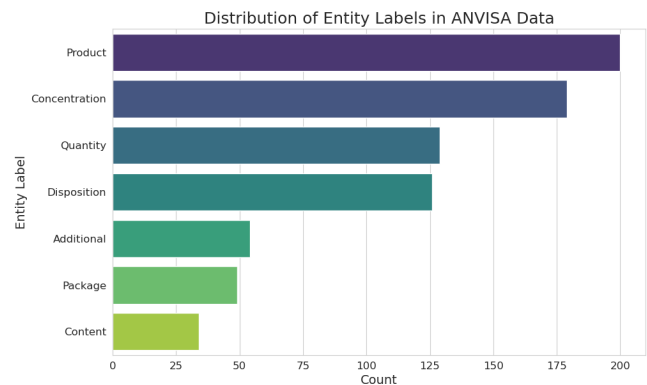


Figure 1. Distribution of labels in the annotated dataset

Our study compares a baseline CRF model with eight variations of BERT-based models for this sequence labeling task. The first approach is based solely on a CRF model, as developed in Darrazão *et al.* [2023] (Section 4.3). The second involves using BERT language models along with classifiers, based on the architectures proposed by Souza *et al.* [2019] (Section 4.4). The CRF model serves as a baseline to assess whether language models can achieve better performance for this task.

For model training, a random sample of product descriptions was selected, annotated, and preprocessed (Section 4.1). Subsequently, a CRF model was constructed to perform the Sequence Labeling task (Section 4.3), using the parameters and ideal features outlined in Darrazão *et al.* [2023]. Eight language models were then trained, varying their architecture (Section 4.5), as detailed in Table 1. These models explore variations in: (1) the base BERT model (Portuguese BERTimbau vs. Multilingual BERT), (2) the training strategy (fine-tuning BERT vs. using BERT as a feature extractor with frozen weights), and (3) the presence or absence of a final CRF layer. Ultimately, each model is evaluated individually and compared with the others (Section 5).

### 4.1 Data Source and Processing

The data for electronic invoice descriptions were provided as part of project 1776 PROMAT/UFAL-SEFAZ-PI, developed by LED/UFAL<sup>1</sup>. The database includes 1 million item descriptions from electronic invoices related to the sale of medications throughout the year 2021 in Brazil, specifically from public procurement contexts.

From the data described above, 200 descriptions were randomly selected, and manual annotations were performed by a domain expert using the Doccano tool<sup>2</sup>. Figure 1 shows the distribution of annotated labels for the dataset.

There are various archetypes for encoding labels used in text annotation tasks, and they directly impact the quality of Sequence Labeling models [Alshammari and Alanazi, 2021]. The BIO scheme was used in all models because the model proposed in Souza *et al.* [2019] is fully compatible with it. Single-word entities were annotated with the B- prefix (e.g., B-PRODUCT). The BIOES scheme was considered but BIO was chosen for consistency with the BERTimbau framework.

In this BIO encoding, prefixes are added for each of the

<sup>1</sup><https://im.ufal.br/laboratorio/led/>

<sup>2</sup><https://doccano.github.io/doccano/>

**Table 1.** Experimental Setup for the Eight BERT-based Language Model Variations. PT refers to BERTimbau (Portuguese), ML to Multilingual BERT. Fine-tuning indicates whether BERT weights were updated. Classifier shows the layers used post-BERT. CRF indicates the presence of a final CRF layer.

Model ID	Corpus-base	Fine Tuning	Classifier	CRF
BERT-CRF PT	BERT Base Portuguese Cased	Yes	Linear	Yes
BERT PT	BERT Base Portuguese Cased	Yes	Linear	No
BERT-BiLSTM-CRF PT	BERT Base Portuguese Cased	No	BiLSTM + Linear	Yes
BERT-BiLSTM PT	BERT Base Portuguese Cased	No	BiLSTM + Linear	No
BERT-CRF ML	BERT Base Multilingual Cased	Yes	Linear	Yes
BERT ML	BERT Base Multilingual Cased	Yes	Linear	No
BERT-BiLSTM-CRF ML	BERT Base Multilingual Cased	No	BiLSTM + Linear	Yes
BERT-BiLSTM ML	BERT Base Multilingual Cased	No	BiLSTM + Linear	No

original 7 categories, namely: B (beginning) for the first word of a named entity; I (inside) for words inside a named entity, which means that whenever the entity consists of more than one token, the first will be labeled as B, and the rest as I; O (outside) for words that do not belong to a named entity.

## 4.2 Performance Evaluation and Experimental Setup

The performance evaluation of the 9 models (1 CRF baseline + 8 BERT variants) is carried out using two tools: during the training and validation of the models, they were evaluated using the SEQeval library<sup>3</sup>, a performance assessment tool for Sequence Labeling with support for different label encoding schemes. The final result (presented here) is obtained using a Perl script<sup>4</sup> provided by the Conference on Computational Natural Language Learning (CoNLL-2000)<sup>5</sup>, a renowned conference in the NLP field that usually holds annual evaluation competitions for tasks in the field. Both SEQeval and the CoNLL script compute entity-level precision, recall, and F1-score based on exact boundary and type matching for recognized entities. It is worth noting that the evaluation metrics of the SEQeval library are equivalent to CoNLL-2000, but the original script was used in the final results to provide further validation for this research.

To enable a fair comparison between all models, the training was carried out using 5-fold cross-validation. The 200 annotated descriptions were divided into 5 folds. In each iteration, 4 folds (80%) were used for training and 1 fold (20%) for validation/testing. The parameters were fixed, and the validation set in each fold was used to evaluate the model trained on the other four folds. The reported results for each model are the average of its performance across these five validation sets. The same subsets were maintained for all models. In other words, for each model presented here, five variants were actually trained, with each of them using one of the training/validation splits derived from the folds. The results for each of these models represent the average of their respective results across the five validation sets.

Therefore, each model will be evaluated using the metrics explained in Section 2.3: Precision, Recall, F1-Score, and Accuracy, as well as the training time and inference time,

both in seconds. Precision, Recall, and F1-score are reported as weighted averages over label types by the evaluation tools.

## 4.3 Sequence Labeling with Conditional Random Fields (CRF)

The model proposed in Darrazão *et al.* [2023] was retrained using the data annotated in the BIO scheme and using the five /validation splits derived from the 5-fold cross-validation constructed earlier (which are used in all the models in this article).

Only minor adjustments were made to the original source code, necessary to train the model with the five distinct sets of training and validation data while storing useful information for evaluation, such as training and inference times. A short script was also created to perform inference on the five respective validation sets with the five models ready and save them in the format expected by the CoNLL-2000 Perl script.

## 4.4 BERT-based Architectures for Sequence Labeling

The BERT-based models in this study adapt the architectures proposed in Souza *et al.* [2019], which combine BERT with a classification layer, optionally followed by a CRF layer. The project utilized the Portuguese language BERT, BERTimbau [Souza *et al.*, 2020a], and also a multilingual BERT variant. Four main architectural variations were implemented, differing in the transfer learning approach (fine-tuning vs. feature-extraction) and the presence of a CRF layer.

The general architecture, depicted in Figure 2, involves the following steps:

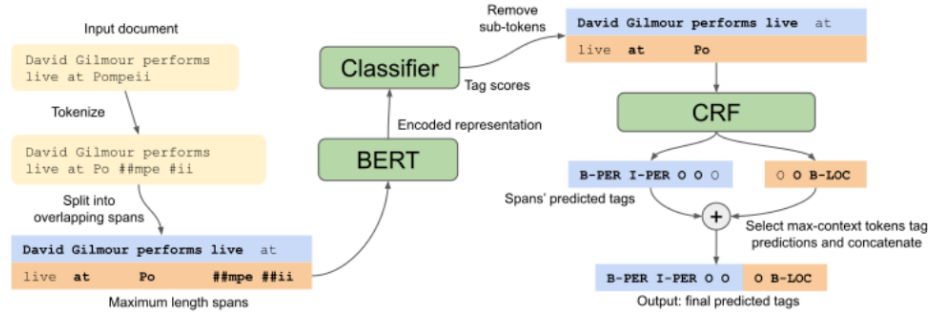
1. **Tokenization and Spanning:** Input descriptions are tokenized using WordPiece [Wu *et al.*, 2016], consistent with BERT’s pre-training. WordPiece splits words into sub-word units representing frequent tokens (e.g., ‘Pompeii’ becomes ‘Po’, ‘##mpe’, ‘##ii’). Long descriptions are divided into overlapping spans of a fixed maximum length (e.g., 128 tokens) with a defined stride (e.g., 64 tokens). This spanning strategy is employed because BERT models have a maximum input sequence length, and it also serves as a form of data augmentation. While it might lead to some context fragmentation at span boundaries, the overlap and the final label selection mechanism (step 6) aim to mitigate this. Combining

<sup>3</sup><https://github.com/chakki-works/seqeval>

<sup>4</sup><https://www.clips.uantwerpen.be/conll2000/chunking/conlleval.txt>

<sup>5</sup><https://www.clips.uantwerpen.be/conll2000/>





**Figure 2.** General architecture for BERT-based sequence labeling models, from Souza *et al.* [2019]. Input text (top-left) is tokenized (e.g., “David Gilmour...”) and split into overlapping spans. BERT generates contextualized embeddings for tokens in each span. These embeddings are passed to a classifier (Linear or BiLSTM+Linear) to predict labels. An optional CRF layer can refine these predictions. The final labels are determined by selecting predictions from spans offering maximum context for each token.

predictions for tokens appearing in multiple spans is handled by the maximum context selection.

2. **BERT Embeddings:** The generated spans are fed into a pre-trained BERT model (either BERTimbau or Multilingual BERT) to produce contextualized word embeddings for each token.
3. **Classification Layer:** These embeddings are passed to a classification layer that maps them to the label space (the 7 entity types + O, in BIO format).

- **Fine-tuning approach:** A simple linear layer acts as the classifier. The weights of BERT and the linear layer are updated during training.
- **Feature-based approach:** The BERT weights are frozen. The classifier consists of a BiLSTM layer followed by a linear layer. The BiLSTM uses a combination of the last four hidden layers of BERT as input features, as suggested by [Devlin *et al.*, 2019].

4. **CRF Layer (Optional):** If included, a CRF layer takes the output scores from the classifier (after removing WordPiece sub-tokens starting with ‘##’) and models label transition probabilities to refine predictions. The CRF uses BERT embeddings and classifier outputs as features. The prediction for each token is made using the Viterbi algorithm in the CRF layer.
5. **Loss Calculation:** Models without a CRF layer are optimized by minimizing cross-entropy loss on the classifier’s output. Models with a CRF layer use the CRF’s loss function.
6. **Label Selection for Overlapping Spans:** For tokens appearing in multiple spans due to overlap, the final predicted label is taken from the span where the token has the maximum context (i.e., is closest to the center of the span).
7. **Final Prediction:** The selected labels for all tokens are concatenated to form the final sequence of predicted labels for the input description.

## 4.5 Training Large Language Models (LLM) for Sequence Labeling

Given that the source code for BERTimbau was publicly available<sup>6</sup>, it was possible to adapt it with minor modifications to allow the training of other models with different datasets and parameters. With it, it is possible to train models with any of the four architectures, using any previously trained BERT model as a starting point: either for fine-tuning or to use its features.

The chosen architectures to be tested (as per Table 1) were selected to allow testing different configurations of using language models for this task. With them, it is possible to contrast the following aspects:

- Which is the more suitable pre-trained model to use as a starting point: BERT trained for Portuguese or BERT trained with 104 different languages;
- What is more effective: fine-tuning the model and adding only a linear classifier layer or keeping the weights of the base model and using them as features for a BiLSTM classifier plus an additional linear layer;
- What is the impact of adding a CRF layer at the end of the architecture?

As for the parameters of the trained models, the same ones proposed in Souza *et al.* [2019] were used. All models used the same seed: 7, per GPU train batch size: 2, gradient accumulation steps: 8, and the number of epochs varied depending on the architecture: BERT-CRF: 15, BERT: 50, BERT-LSTM-CRF: 50, BERT-LSTM: 100. It is worth noting that the models were trained using a GPU with 12GB of dedicated memory available.

## 5 Results and Discussion

Table 2 shows the average results of all experiments across the 5 cross-validation folds. In the first column is the model identification, which can be cross-referenced with Table 1, since both are in the same order (besides the first one, which is the base model). Names wise, all of them are BERT based, except the first. If there is a BiLSTM in the name, this is in the classifier layer and thus has no fine-tuning; if not, it is just a

<sup>6</sup><https://github.com/neuralmind-ai/portuguese-bert>

**Table 2.** Results Table for Sequence Labeling Models (with standard deviations for LLM-based models)

	Model	F1-Score	Accuracy	Precision	Recall	Training (sec)	Inference (sec)
1	CRF Base	78.17	76.85	<b>84.81</b>	73.12	<b>4.66</b>	<b>0.0001</b>
2	BERT-CRF PT	<b>82.19</b> $\pm$ 5.08	<b>82.39</b> $\pm$ 5.18	84.40 $\pm$ 5.25	80.23 $\pm$ 6.22	119.34 $\pm$ 1.43	1.04 $\pm$ 0.01
3	BERT PT	78.84 $\pm$ 5.33	78.31 $\pm$ 4.93	77.87 $\pm$ 7.41	80.11 $\pm$ 5.41	567.91 $\pm$ 27.69	1.11 $\pm$ 0.21
4	BERT-BiLSTM-CRF PT	79.13 $\pm$ 2.38	79.13 $\pm$ 3.06	83.17 $\pm$ 5.49	75.76 $\pm$ 3.81	108.83 $\pm$ 3.63	1.08 $\pm$ 0.05
5	BERT-BiLSTM PT	80.68 $\pm$ 6.36	80.00 $\pm$ 6.65	79.35 $\pm$ 8.16	82.17 $\pm$ 4.52	54.90 $\pm$ 0.94	1.00 $\pm$ 0.06
6	BERT-CRF ML	77.52 $\pm$ 5.40	77.66 $\pm$ 3.66	82.00 $\pm$ 5.83	73.54 $\pm$ 5.36	138.81 $\pm$ 2.45	1.87 $\pm$ 0.04
7	BERT ML	78.86 $\pm$ 7.74	78.09 $\pm$ 7.69	77.78 $\pm$ 10.73	80.25 $\pm$ 4.80	419.94 $\pm$ 4.62	1.84 $\pm$ 0.02
8	BERT-BiLSTM-CRF ML	80.02 $\pm$ 1.92	79.94 $\pm$ 1.99	83.86 $\pm$ 5.09	76.80 $\pm$ 3.70	117.05 $\pm$ 2.76	1.85 $\pm$ 0.03
9	BERT-BiLSTM ML	80.64 $\pm$ 5.53	80.03 $\pm$ 5.19	78.68 $\pm$ 5.81	<b>82.78</b> $\pm$ 5.94	64.28 $\pm$ 1.19	1.74 $\pm$ 0.05

linear layer, and the model was finetuned. If there is a CRF in the name, there is a CRF layer in the model. The last two characters represent the pretrained model used, PT for Portuguese and ML for Multilingual. The other columns represent the model’s respective average metrics, as explained in Section 4.2. Standard deviations are provided for the LLM-based models (these statistics were not collected for the baseline model). The standard deviations are added to provide a sense of variation in the runs, not to statistically prove or disprove hypothesis (such experiment would require an extensive design of representative datasets and tasks that are outside the scope of this paper).

Beginning with the main question: can language models be effectively used to improve the performance of the sequence labeling task on descriptions of medical products in invoices? The answer appears to be affirmative, with some caveats.

It is noticeable that applying an approach based on LLMs generally improves the results, as 7 out of 8 LLM models achieve a better F1-Score than the Base CRF, reaching an improvement of approximately 4% in the best case (BERT-CRF using a pretrained Portuguese model). However, the improvements are modest, and a statistical significance test would be needed to confirm if these differences are not due to chance, especially given the dataset size and the number of models compared.

With the questions raised in Section 4.5, we can analyze the components used in these models:

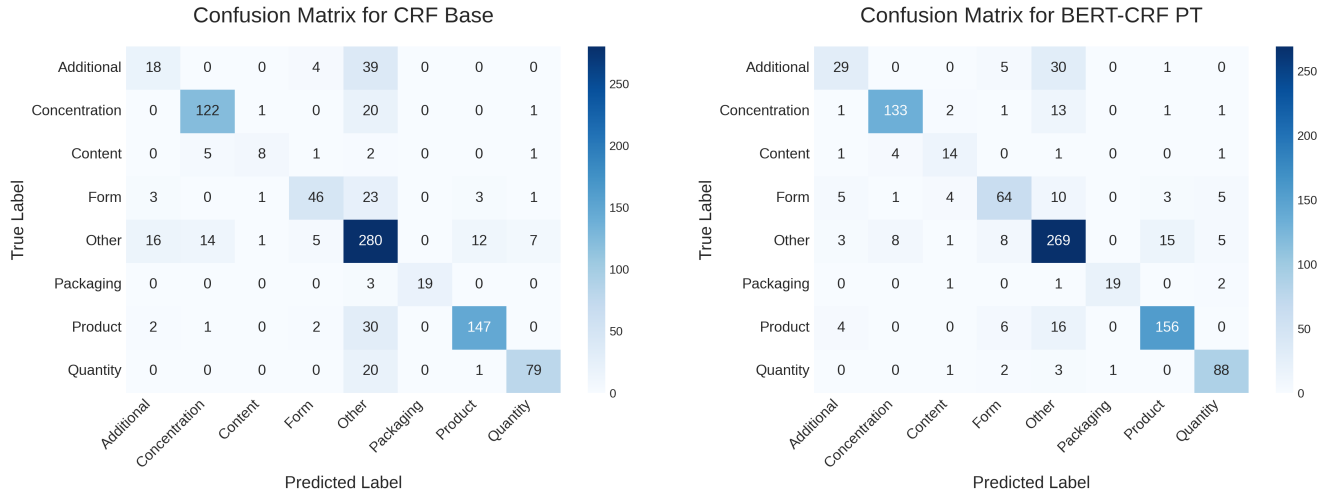
- Which is the more suitable pre-trained model to use as a starting point: BERT trained for Portuguese or BERT trained with 104 different languages?
  - The difference in results when varying the pre-trained models (Portuguese and Multilingual) is generally small, except in the case of BERT-CRF. Not only does the BERT-CRF PT model achieve the best results, but using Portuguese as a starting point also leads to an improvement in F1-Score (an increase of 4.67% compared to the BERT-CRF PT. Comparing mean performance of all PT models (Mean F1 = 80.21) vs. all ML models (Mean F1 = 79.26), the Portuguese models show a slight edge on average.
- What is more effective: fine-tuning the model and adding only a linear classifier layer or keeping the weights of the base model and using them as features for a BiLSTM classifier plus an additional linear layer?

- Overall, a feature-based approach with BiLSTM yields comparable or slightly better results in most configurations. However, the highest performing model overall (BERT-CRF PT) utilized the fine-tuning approach.

- What is the impact of adding a CRF layer at the end of the architecture?
  - The result here is mixed. In most cases, not using a CRF layer at the end resulted in slightly better or comparable F1-scores. However, the overall best-performing model (BERT-CRF PT) included a CRF layer. This suggests the CRF’s benefit might be configuration-dependent and particularly synergistic with the fine-tuned Portuguese BERT.

The outcome here is that there appears to be a good synergy of specific components, specifically in the architecture that was trained using a pre-trained Portuguese BERT while fine-tuning and using a final CRF layer. Those components alone do not consistently guarantee improvement: choosing between the two different pretrained models in most cases does not change much. The fine-tuning approach alone is not always better than using the pretrained model weights with a BiLSTM. Adding a final CRF layer often slightly lowers the F1-Score of the models. But when we combine BERTimbau, fine-tuning, and a CRF layer, we get the best result among all of them. The distinct behavior of the BERT-CRF PT model, particularly its positive response to fine-tuning and CRF, might suggest it’s an outlier or that these components interact uniquely well for this specific base model and task; inspecting individual fold results might offer more insight.

Another analysis we can perform is that, even though F1-Score is the most commonly used metric for model evaluation (as it combines precision and recall), in cases where one of the two is more important than their combination, the other approaches may be more suitable. For example, the Base CRF achieved the best precision among all models, while the BERT-BiLSTM ML architecture achieved the best recall. Furthermore, there is the issue of training and inference times, where, using the Base CRF, both times are significantly lower than any of the approaches using LLMs. It is worth noting that training was conducted using only 200 descriptions; in a large-scale application, significantly more time and processing power would be required. Figure 3 shows the confusion matrices for the *Base CRF* and the *BERT-CRF PT* models for a more detailed comparison of model performance. The



**Figure 3.** Confusion matrices for Base CRF and best BERT performing model.

figure shows similar patterns in the errors, with a tendency of the baseline model to categorize tokens as *other/outside*.

We can argue that the ideal model depends heavily on the use case. The variables to consider when choosing the model include the main objective (i.e. which metric is more important for the application), the complexity, the available time for training and inference, and the available computational power. The Base CRF model can be easily trained using only CPU and RAM, while LLM models become impractical if you do not have a GPU to assist with training.

Lastly, complexity is a very important point, considering the difficulty and learning curve to use different approaches. When working with a basic CRF model, each corpus requires a very specific set of features, which is obtained through intensive analytical study and experimental tests, as can be seen in Darrazão *et al.* [2023]. Each dataset has its own characteristics, making it difficult to reuse already designed basic CRF models, resulting in new planning and design every time it is necessary to apply such a model to a new database.

On the other hand, using LLM models like those based on the BERTimbau framework [Souza *et al.*, 2019], requires an initial investment in understanding how the model works and how to properly adapt the data, as well as to ensure that the results obtained make sense and are correct. But once the architecture is ready and defined, as in this case, only small changes and data treatments are needed to fit the data into the model, and then the model is ready to be used.

## 6 Conclusion

This article evaluated the performance of different models for the task of sequence labeling in Brazilian invoice descriptions, comparing a basic CRF model with transformer-based models. When comparing the models, it is possible to state that using Large Language Models (LLMs) for sequence labeling within the proposed scope can yield better results in many, though not all, scenarios evaluated, with the best LLM configuration outperforming the CRF baseline. However, the magnitude of these improvements is generally modest, and careful consideration of trade-offs is necessary.

The use of LLMs in this scenario should not be taken lightly; it is necessary to analyze the use case to determine the most appropriate approach. Although they can deliver better results, they require more resources for training and utilization. If available, in conjunction with the provision of a pre-structured model and a professional who knows how to use it, making it ready for training and use, regardless of the corpus, is relatively straightforward. In contrast, training a basic CRF model, despite having faster training and inference times, requires significantly more time for development and in this study, yielded lower F1-scores compared to the best LLM.

The amount of data used for training and evaluating those models was considerably low (200 descriptions). It is possible that increasing the volume of data could lead to a larger discrepancy in performance between the LLM-based models and the CRF Basic model. LLMs have the capability to potentially learn more complex patterns from the data compared to the features defined in the CRF Basic model. Consequently, increasing the quantity of available data might further amplify the dissimilarity between the descriptions. The experiment employed only one expert for the text annotations due to time and budget constraints. While it is unlikely that the personal bias would benefit specifically one type of approach and change significantly the results, more annotators and an agreement analysis would be beneficial to the quality of the models.

Thus, the results obtained and discussed here can serve as a basis for various works that require structuring the content of fiscal documents and can be used to perform similar analyses on other datasets where sequence labeling is desired.

### 6.1 Future Work

The test set used is limited and does not reflect the volume and diversity of a production environment. Future tests could take into account larger datasets and performance requirements, as augmenting the volume of data during training and evaluation might result in more pronounced differences in model evaluation.



Reconsidering the granularity of data labels is advisable, particularly regarding words referring to pharmaceutical laboratories or indicating whether it is a generic product, which were labeled as ‘Additional’. Revisiting these labels has the potential to enhance the performance of the models.

Finally, it would be interesting to test models that rely on Transformers but with different approaches, like GPT that uses a decoder-only structure (instead of encoder-only like BERT) which may bring new findings and achieve even better results, using methods like few-shot learning or even zero-shot learning. Additionally, exploring more recent and larger LLMs, if resources permit, could be beneficial.

## Declarations

### Authors’ Contributions

Gomes-Jr and Oliveira contributed to the conception of this study and supervised Darrazão in performing and reporting the experiments. Darrazão is the main contributor and writer of this manuscript. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### Availability of data and materials

The datasets generated and/or analyzed for the study were kept private due to the confidential nature of the end task of the public institutions involved.

## References

- Alshammari, N. and Alanazi, S. (2021). The impact of using different annotation schemes on named entity recognition. *Egyptian Informatics Journal*, 22(3):295–302. DOI: 10.1016/j.eij.2020.10.004.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623. DOI: 10.1145/3442188.3445922.
- Bose, P., Srinivasan, S., Sleeman, W. C., Palta, J., Kapoor, R., and Ghosh, P. (2021). A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319. DOI: 10.3390/app11188319.
- Cao, P., Wang, Y., Zhang, Q., and Meng, Z. (2023). GenKIE: Robust generative multimodal document key information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14620–14631. Association for Computational Linguistics. Available at: <https://aclanthology.org/2023.findings-emnlp.979>.
- Darrazão, E., Amorim, V., Oliveira, K., and Gomes-Jr, L. (2023). Engenharia e avaliação de features para extração de informação em notas fiscais. pages 80–89. DOI: 10.5753/erbd.2023.229441.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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, Minneapolis, Minnesota. Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423.
- F. N. de Oliveira, L. P. G. d. S. (2020). Estratégias para combater a sonegação fiscal: Um modelo para o icms baseado em redes neurais artificiais. *Revista de Gestão, Finanças e Contabilidade*, 10:42–64. DOI: 10.18028/rgfc.v10i1.7474.
- Ganiz, M. C., Celik, M., Celikmasat, G., Aydin, G., and Yuret, D. (2022). Biomedical named entity recognition using transformers with bilstm+crf and graph convolutional neural networks. In *2022 16th International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–6. IEEE. DOI: 10.1109/INISTA55331.2022.9872223.
- He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., and Jiang, S. (2020). A survey on recent advances in sequence labeling from deep learning models. DOI: 10.48550/arxiv.2011.06727.
- Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
- Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*. Available at: <https://arxiv.org/abs/1508.01991>.
- John Lafferty, Andrew McCallum, F. C. P. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. ICML*. Available at: [https://repository.upenn.edu/cis\\_papers/159/](https://repository.upenn.edu/cis_papers/159/).
- Jurafsky, D. and Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. Prentice Hall. DOI: 10.1162/coli.B09-001.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. DOI: 10.48550/arXiv.1301.3781.
- Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. DOI: 10.3115/v1/D14-1162.
- Pereira, R. d. S. (2020). Redes heterogêneas para classificação de produtos em notas fiscais eletrônicas de compras públicas [tcc]. *CGU*. Available at: <https://repositorio.cgu.gov.br/handle/1/64722>.
- Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training. Available at: <https://api.semanticscholar.org/CorpusID:49313245>.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D.,

- and Sutskever, I. (2019). Language models are unsupervised multitask learners. Available at: <https://api.semanticscholar.org/CorpusID:160025533>.
- Seymore, K. and Rosenfeld, R. (1999). Learning hidden markov model structure for information extraction. Available at: <https://www.cs.cmu.edu/~roni/papers/iestruct-aaaiws99.pdf>.
- Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*. DOI: 10.48550/arxiv.1909.10649.
- Souza, F., Nogueira, R., and Lotufo, R. (2020a). Bertimbau: Pretrained bert models for brazilian portuguese. In Cerri, R. and Prati, R. C., editors, *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing. DOI: 10.1007/978-3-030-61377-8\_28.
- Souza, F., Nogueira, R., and Lotufo, R. (2020b). Portuguese named entity recognition using BERT-CRF. In *Proceedings of the 14th International Conference on the Computational Processing of Portuguese (PROPOR)*, pages 304–313. Springer. Preprint available as arXiv:1909.10649 [cs.CL]. DOI: 10.48550/arxiv.1909.10649.
- Sutton, C., McCallum, A., et al. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373. DOI: 10.48550/arXiv.1011.4088.
- Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proc. Conference on Natural Language Learning at HLT-NAACL*. Available at: <https://aclanthology.org/W03-0419>.
- Tourille, J., Doutreligne, M., Ferret, O., Névél, A., Paris, N., and Tannier, X. (2018). Evaluation of a sequence tagging tool for biomedical texts. In *Proc. International Workshop on Health Text Mining and Information Analysis*. DOI: 10.18653/v1/W18-5622.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. DOI: 10.48550/arXiv.1706.03762.
- Veras Carvalho Menezes, A. P. (2022). Inteligência artificial para identificação de indícios de fraude e corrupção em compras públicas no tcu. *Revista Debates em Administração Pública – REDAP*, 3(2). Available at: <https://www.portaldeperiodicos.idp.edu.br/redap/article/view/6521>.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. DOI: 10.48550/arXiv.1609.08144.