Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Combining prompt-based language models and weak supervision for named entity recognition from legal documents

Vitor V. Oliveira

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Thiago de Paulo Faleiros

Coorientador
Prof. Dr. Ricardo Marcondes Marcacini

Brasília
2023

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

# Combining prompt-based language models and weak supervision for named entity recognition from legal documents

Vitor V. Oliveira

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Thiago de Paulo Faleiros (Orientador)
CIC/UnB

Prof. Dr. Guilherme Novaes Ramos
CIC/UnB

Prof. Dr. Marcelo Ladeira
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 07 de fevereiro de 2023

# Dedication

I dedicate this work to all my family, friends, and my girlfriend, who supported me at all times.

# Acknowledgements

I thank the faculty of the Departments of Computer Science and Mathematics at the University of Brasília for all the teachings and contributions to my education. I am grateful to the *Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF)* for the financial incentives that allowed me to participate in numerous educational and research projects throughout my undergraduate studies. I would like to thank my academic supervisor Thiago de Paulo Faleiros and my academic co-supervisor Ricardo Marcondes Marcacini, for all their understanding, welcome and support; and to all the collaborators who took part in this work for their disposition and teachings.

This work was carried out with the support of the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES)*, through the Access to the Periodicals Portal.

# Resumo

O Reconhecimento de Entidades Nomeadas (NER) é uma tarefa muito relevante para a recuperação de informações textuais em problemas de Processamento de Linguagem Natural (NLP). O estado da arte dos métodos de NER mais recentes exigem que humanos anotem e forneçam dados para o treinamento de modelos de aprendizado profundo. No entanto, usar força humana para identificar, circunscrever e rotular entidades manualmente pode ser extremamente caro em termos de tempo, dinheiro e esforço. Este artigo investiga o uso de modelos de linguagem baseados em prompt (OpenAI's GPT-3) e supervisão fraca para a rotulação de textos de domínio jurídico. Aplicamos ambas estratégias como abordagens alternativas ao método tradicional de anotação baseado em força humana, contando com poder computacional em vez de esforço humano para rotular dados textuais, subsequentemente, comparamos os desempenhos de modelos gerados por computadores e modelos gerados por humanos. Também introduzimos combinações de todos os três métodos mencionados (modelos baseado em prompt, supervisão fraca e anotação humana), com o objetivo de encontrar maneiras de manter alta eficiência e baixo custo de anotação. Mostramos que, apesar da rotulação humana ainda manter melhores resultados de desempenho geral, as estratégias alternativas e suas combinações se apresentaram como opções válidas, exibindo resultados positivos e performance semelhantes a custos mais baixos. Resultados finais demonstram uma preservação de desempenho médio em relação a rotulação humana de 74,0% para o GPT-3, 95,6% para a supervisão fraca, 90,7% para a combinação de GPT + supervisão fraca e 83,9% para a combinação de GPT + 30% rotulação humana.

**Palavras-chave:** Reconhecimento de Entidades Nomeadas, OpenAI's GPT-3, Supervisão Fraca

# Abstract

Named Entity Recognition (NER) is a very relevant task for text information retrieval in Natural Language Processing (NLP) problems. Most recent state-of-the-art NER methods require humans to annotate and provide useful data for model training. However, using human power to identify, circumscribe and label entities manually can be very expensive in terms of time, money, and effort. This paper investigates the use of prompt-based language models (OpenAI's GPT-3) and weak supervision in the legal domain. We apply both strategies as alternative approaches to the traditional human-based annotation method, relying on computer power instead human effort for labeling, and subsequently compare model performance between computer and human-generated data. We also introduce combinations of all three mentioned methods (prompt-based, weak supervision, and human annotation), aiming to find ways to maintain high model efficiency and low annotation costs. We showed that, despite human labeling still maintaining better overall performance results, the alternative strategies and their combinations presented themselves as valid options, displaying positive results and similar model scores at lower costs. Final results demonstrate preservation score of human-data trained models averaging 74.0% for GPT-3, 95.6% for weak supervision, 90.7% for GPT + weak supervision combination, and 83.9% for GPT + 30%Human-labeling combination.

**Keywords:** Named Entity Recognition, OpenAI's GPT-3, Weak Supervision.

# Contents

# List of Figures

# List of Tables

# Acronyms

**CD** Critical Difference.

**DODF** Official Gazette of the Federal District.

**GPT** Open AI's Generative Pre-Training Transformer.

**GPT-3** Open AI's Generative Pre-Training Transformer 3.

**HMM** Hidden Markov Model.

**IOB** Inside–Outside–Beginning token format.

**NER** Named Entity Recognition.

**NLP** Natural Language Processing.

# Chapter 1

# Introduction

The Official Gazette of the Federal District (*Diário Oficial do Distrito Federal (DODF)*) is a publication that contains reports on Brazilian government actions and is updated daily. The DODF includes information about retirements, public procurement, decrees, and other matters, organized by the government agency. The documents in each edition of the DODF are called acts and are divided into different types. These types are grouped into sections: Section I includes normative acts of general interest, such as laws, decrees, and resolutions; Section II contains information related to civil servants; and Section III covers bidding processes and information about contracts.

The Official Gazette of the Federal District (DODF) can be a valuable resource for anyone or any organization that wishes to monitor government activities. Though it serves as a transparency journal, it is mainly used by public officials and other professionals for professional purposes. The DODF can be used to verify if something has been made official and to access information about an action, such as the date or the agency involved. It allows tracking information such as companies hired by the government, the career progression of civil servants, and much more. However, extracting information from the DODF by Natural Language Processing can be challenging because no detailed information in the acts is labeled.

Natural Language Processing (NLP) is a vast branch of computer science and artificial intelligence that can be defined as a series of computerized approaches and techniques for analyzing and representing naturally occurring texts [1, 2]. One of NLP's most acknowledged and challenging tasks is Named Entity Recognition (NER) which consists in identifying and classifying specific types of information elements, called named entities, in natural texts [3]. In the NER task, given a sample text, key named entities, such as "Name", "Place" or "Value", are defined and the machine must efficiently identify them by using several differentiation factors, for example, "Name" type entities always presents their initial letter capitalized.

Regarding information extraction in legal domain documents, such as DODF, NER is one of the main options to retrieve knowledge efficiently [4]. Legal Texts are usually lengthy and complex by nature, preferring formal structure rather than readability. However, legal bodies' language cores contain large collections of patterns and identifiers that can define many rules for NER classifiers and further enable their use [5]. In most recent years, state-of-the-art computation techniques used for the NER task normally include training and testing Deep Neural Network models [6]. Although such models obtain promising results, they normally depend on large pools of data to provide a reliable source of information for training and achieving ideal behavior [7].

In this context, there is a growing need of dependable labeled data to provide the models, which commonly is supplied by human effort, by manually annotating and categorizing texts. Even though this human participation improves model performance, in many projects, the usual process of reading, searching, identifying, circumscribing, and reviewing can be costly in terms of time, money, and effort. Our paper aims to explore alternative strategies to this traditional NER approach, introducing solutions using weak labeling, machine learning, and other computational methods to target cost and effort reduction.

In particular, we focus on two recent and promising strategies for dealing with tasks involving unlabeled data. The first is the use of prompt-based language models, such as OpenAI GPT-3, to leverage pre-trained models to generate texts with predefined structures and extract relevant information for a given domain. The second consists of weak supervision techniques incorporating rule-based systems and heuristics to generate labeled datasets quickly. Below, we present a brief description and motivation of these strategies to support the data labeling of legal documents.

Generative Pre-trained Transformer 3, or GPT-3, is an autoregressive language model released by the non-profit research organization OpenAI [8]. GPT-3 makes use of deep learning techniques to generate, from prompt-based inputs, natural language texts that are remarkably hard to distinguish from human-authored content. The transformer model has proved to be very potent and efficient in many NLP tasks, even though it shows limitations in semantic coherence and generation of unreal assertions [9]. The proposal here is using GPT-3 abilities to address the NER task, generating texts to actively predict named entities of a given instance and efficiently supply labeled data for training models.

Weak Supervision or Weak Labeling is another alternative to support data labeling since it focuses on bootstrapping new labeled data with computer effort [10]. It consists of combining noisy, limited, or imprecise sources of supervision signals, such as rule-based systems or other machine learning models, to obtain probabilistic labels for large

amounts of unlabeled data [11, 12][1]. In better terms, weak supervision works by heuristically creating its own categorized data, relying on label functions to annotate documents automatically. These functions can use different strategies in their labeling processes such as regular expression patterns, class-indicative keywords, or heuristic methods [13, 14].

The main contributions of this paper are:

- Prompt-based language models (GPT-3) and Weak Supervision in NER: Study and analysis of GPT-3 and weak supervision uses in Named Entity Recognition tasks for legal domain, as well as its capacity to efficiently reduce annotation cost and maintain model accuracy.
- Exploring Combinations: Exploring combinations of the previously mentioned techniques and traditional human labeling, aiming to achieve better model accuracy while maintaining low-cost annotations.

The remaining of this work is structured as follows. In Section 2, we present the most related works to our proposal, that we could find in the literature. In Section 3, we present an overview of our research methodology focusing on dataset characteristics, the two previously mentioned labeling approaches(prompt-based models and weak supervision), and the four chosen models and their training hyperparameters. In Section 4, we present the experimental evaluation we performed and the results we obtained. Finally, in Section 5 we draw our main conclusions from this work and present some perspectives on future works.

[1] `https://ai.stanford.edu/blog/weak-supervision/`

# Chapter 2

# Related Work

The first work to explicitly define NER as a term in the legal domain was Dozier et al. [15]. In this paper, named entity recognition and resolution in legal documents are thoroughly discussed, and afterward examined by the implementation of named entities lists lookups, contextual rules, and statistical models applied in the US case law and many other legal documents. By using the three mentioned methods, the paper then describes an actual system capable of efficiently identifying named entities in legal texts and subsequently evaluates its accuracy.

Further works such as Vardhan et al.[16] display NER as a powerful method to correctly recognize numerous entities and relevant information in legal bodies. The paper also approximates itself to state-of-the-art techniques, using deep neural network elements, such as convolutional neural networks and multi-layer perceptrons, to build an effective NER model for legal information extraction.

Regarding NER uses in Brazilian legal documents, also our paper's main language domain, Luz de Araujo et al. [17] presents LeNER-Br, a dataset specifically constructed for this task. After the dataset confection, the paper proceeds on training a relevant state-of-the-art machine learning model, LSTM-CRF, and achieves good performance averaging F1 scores of 92.53%, verifying the viability of the proposed dataset and NER models for legal applications.

Related to the proposed GPT-3 approach there are many works that stand out. Firstly, Wang et al. [18] explores ways to leverage GPT-3 as a low-cost data labeler for model training. Results here show that GPT-3 can achieve the same model performance in a variety of Natural Language Understanding and Natural Language Generation tasks as human labels while maintaining a cost reduction of 50% to 96%. Furthermore, it is also proposed a novel framework of combining pseudo labels from GPT-3 with human labels, which leads to even better performance with a limited labeling budget.

Meyer et al. [19] propose approaches such as synthetically generating data and data

augmentation using prompt-based GPT-3. They investigate the feasibility and cost-benefit trade-offs of using non-fine-tuned synthetic data to train classification algorithms, comparing performance between classifiers trained with synthetically generated data and real user data. The conclusion shows that, although the trained classifiers perform much better than random baselines, their performance does not compare to classifiers trained on even small amounts of real user data, largely due to lacking variability. Also, it is concluded that synthetically generated data might be preferable to the collection and annotation of naturalistic data.

Given GPT-3 capabilities and considerate power in text generation, many questions regarding its benefits and hazards have been raised. The article by Floridi and Chiriatti et al. [20] expands a thorough analysis of GPT-3's scope and nature, specifically focusing on experimenting with its mathematical, semantic, and ethical aspects. The paper shows that while GPT-3 is indeed a powerful technology, it still failed all three experiments, not always being able to solve simple mathematical expressions, presenting no understanding of semantics and contexts, and reflecting humanity's worst tendencies and ethical issues such as racism. Still, even though GPT is not an intelligent or sensible AI, its remarkable ability for generating well-structured and syntactically coherent texts is unquestionable, presenting many applications for the right tasks.

Now considering weak supervision, there is also a broad spectrum of works in the area. For example, Ratner et al. [21] presents *Snorkel*, a system that uses weak supervision to train state-of-the-art models without hand labeling any training data. *Snorkel* delivers a flexible interface, allowing users to efficiently write labeling functions that express arbitrary noisy heuristics, afterward applying and denoising these functions in an end-to-end implementation. It is shown that *Snorkel* enables building models 2.8× faster and increases predictive performance by an average of 45.5% versus hand labeling.

Similar to Ratner et al. [21] approach, Karamanolakis et al. [22] develops *ASTRA*, a framework for iterative self-training of deep neural networks with weak supervision, aiming to improve on the effectiveness of regular weak supervision of frameworks like *Snorkel*. Results show that the strategy implemented, using self-training models and rule attention, presented significant improvements over state-of-the-art baselines.

Probably the most aligned weak supervision inspiration to our article is Lison et al. [23]. In this work, weak supervision is presented in a broad spectrum of labeling functions specifically targeting the NER task. Also, the paper follows an approach based on label function aggregation using a Hidden Markov Model (HMM), which will also be used in our paper, for capturing functions with varying accuracies and mislabeling. Essentially, a successful NER approach is achieved, well suited for sequence labeling tasks and probabilistic labeling predictions by utilizing the aggregation of a vast number of

different label functions.

# Chapter 3

# Research Methodology

In this paper, we investigated prompt-based language models and weak supervision to address the drawback of high labeling cost and annotation effort in legal domain texts. An overview of the research methodology proposed in this paper is illustrated in Figure 3.1, involving four steps: data gathering, prompt-based language model labeling, weak supervision labeling, and models and training. This section will discuss in detail the procedures and techniques adopted by each one of these four steps.

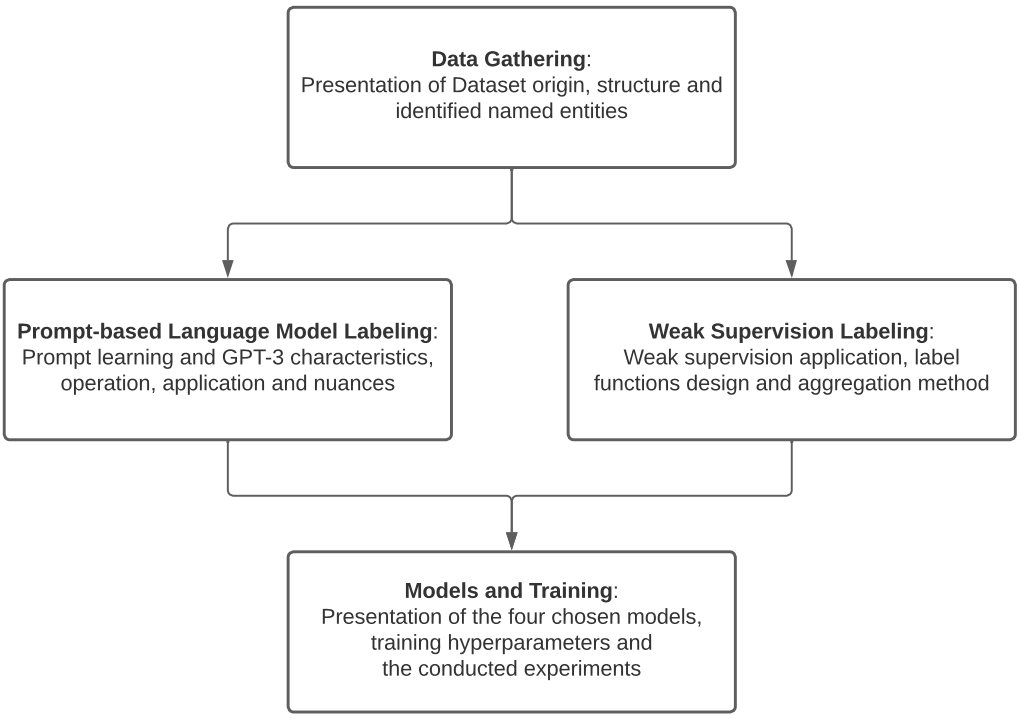


Figure 3.1: Methodology workflow.

## 3.1 Data Gathering

The Dataset used for this experiment was the *DODFCorpus I: Atos de Contratos e Licitações*, provided by the project UnB-KnEDLe from *Universidade de Brasília*. It consists of a Brazilian Portuguese dataset extracted from the previously mentioned Official Gazette of the Federal District (*Diário Oficial do Distrito Federal (DODF)*), a daily updated public document from the Brazilian capital Brasília and its federation unit the *Distrito Federal*. DODF reports on Brazilian government actions and contains all the acts of public administration and services conducted in the region. We will be strictly focusing on the 1.542 instances composing the "Contract" acts, a specific type of act in the corpus that corresponds to a regulated contract between companies and the public state.

We partition the dataset via random sampling, and considering GPT-3 prompt-size limitations, into 783 training acts, 379 validation acts a 380 testing acts, as can be seen in the superior part of Table 3.1. Also, ¨Table's 3.1 inferior part displays a listing of all extracted named entities and their description. Lastly, figure 3.2 presents a practical "Contract" labeling example.

Table 3.1: DODFCorpus dataset partitions and "Contract" named entities.

| **Training** | **Validation** | **Testing** |
|---|---|---|
| 783 | 379 | 380 |

| **Named Entities** | **Entity Description** |
|---|---|
| contract_number | Contract identification number. |
| GDF_process | Process number before the Federal District government (GDF). |
| contractual_parties | Combination of contracting body, contracted entity, and convening entities. |
| contract_object | Object to which the contract refers. |
| contract_date | Contract signature date. |
| contract_value | Estimated contract final value. |
| contract_duration | Contract term of validity. |
| budget_unit | Contract budget union number. |
| work_program | Contract work program number. |
| nature_of_expenditure | Contract nature of expenses number. |
| commitment_note | Contract commitment note. |

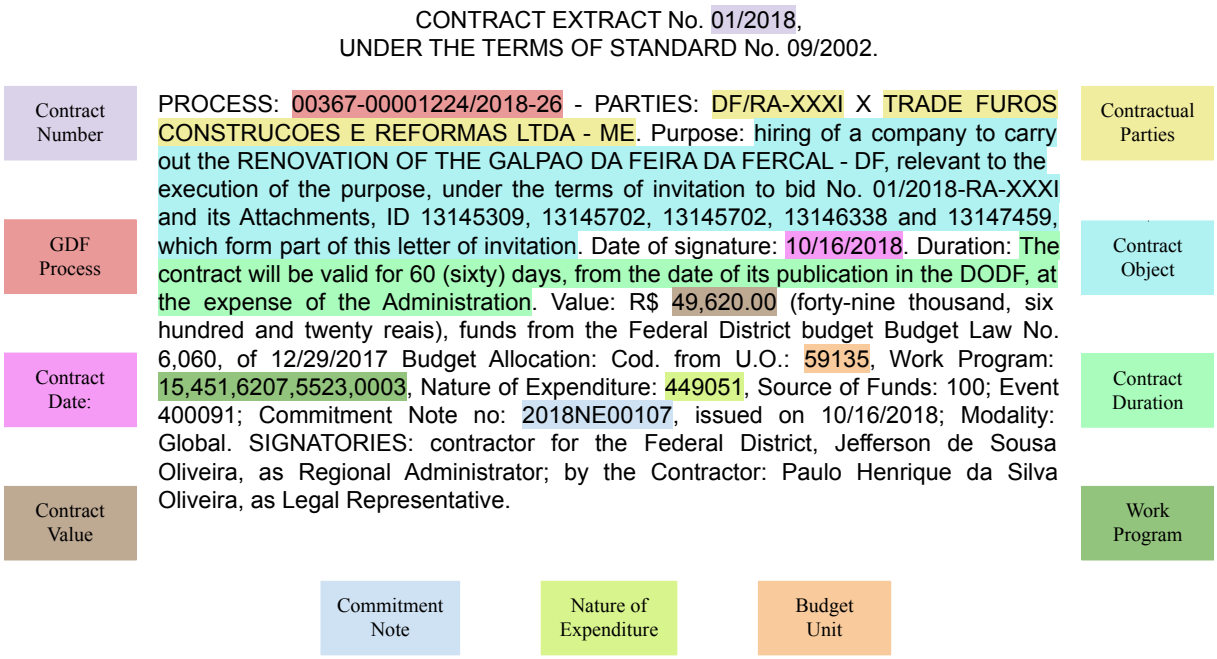


Figure 3.2: “Contract” act labeling example.

## 3.2 Prompt-based language model labeling

In this section, we propose the use of language models trained on large datasets to support data labeling. The most recent models use the prompt learning mechanism, in which the task definition can be described in the data input and formulated as a masked language modeling task. An advantage of the prompt-learning mechanism is the reuse of language models without the need for retraining, which is potentially useful for transfer learning and few-shot learning [24]. Here, we exploit prompt learning to extract entities from new texts with a few annotated examples from legal documents.

The GPT models (versions 2 and 3) stood out with the use of prompt learning. In this work, we investigate the recent GPT-3 [8] model which is composed of four main models with different levels of power suitable for different tasks. The model used in this experiment was Davinci, GPT-3’s most capable model and also most expensive, having a cost of $0.02 for 1,000 GPT-3 tokens, which corresponds to an average of 750 words.

The Davinci model used in this paper is able to compute only a maximum request of about 2,049 tokens for context (prompt and completion). Given the fact that some “Contracts” instances presents a considerably large size, not all of the 1.542 “Contracts” acts fit those descriptions. With that in mind, GPT-3 predictions were only successfully applied to 783 acts of the complete dataset, establishing this acts as the final training dataset and hereafter splitting the 759 remaining acts into 380 for testing and 379 for

validation.

Thereby, the complete GPT-3 labeling process consists in prompting Davinci examples of labeled data and collecting the predictions made in the 783 acts of the training base. For this, three dataset instances were handpicked according to their structures and label occurrences. Then, these instances were selected one at a time, in a randomly and evenly distributed manner, and given as prompts to GPT-3's Davinci model, which next applies its prediction method to exactly one unlabeled act for each selection, based in the selected prompt example. This process ensures the occurrence of all of the eleven "Contracts" named entities, since there are missing entities in some acts, and ensure GPT-3's adaptability to different structured acts. Finally, by the end of these procedures, the 783 training instances were successfully labeled by GPT-3, which applied its predictions in a total of 1.565.108 tokens and had a final cost of $31.30 dollars.

Next, in figure 3.3 we present a practical example referent to GPT-3's prompt-labeling process mentioned in the previous paragraph, using a "Contract" text as input and showing the expected output. In the figure, we can observe how GPT-3 uses the prompt as a labeling example, generating an annotation text output of the unlabeled "Contract" that resembles the output example given by the prompt.

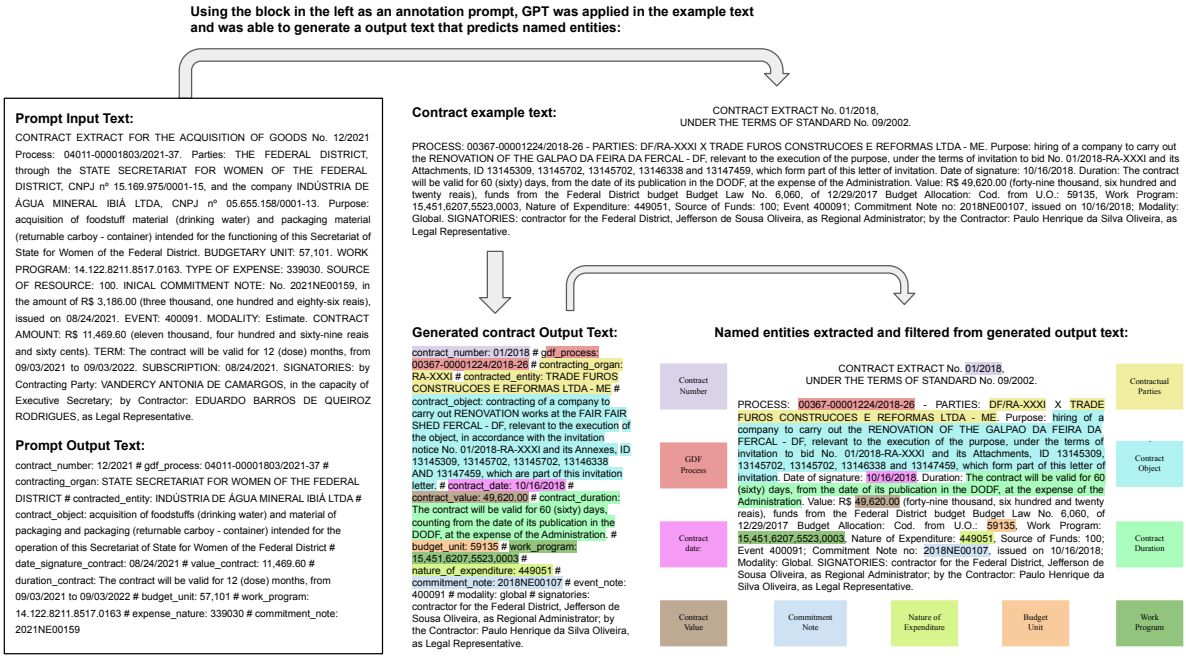


Figure 3.3: GPT-3 prompt-labeling process example.

## 3.3 Weak Supervision Labeling

Regarding the weak supervision labeling process, the approach consists in developing label functions capable of efficiently identifying named entities focusing on different characteristics, strategies, and heuristics. Afterward, label function results must be denoised and merged using an effective aggregation method, finally generating a more accurate and refined label prediction.

We designed two types of label functions for each entity of the "Contract" dataset. The first one is a regular expression (regex) oriented label function, it uses regex to specify and identify a search pattern in the act's text. The second is a keyword detection-oriented function, that is, it uses the occurrence of specific words, punctuation, and symbols to establish starting and ending patterns for all entities. Both types of labeling functions were designed and applied for each entity present in the dataset and had their results combined with one another by the use of the Hidden Markov Model [25] aggregation method.

All weak supervision labeling functions were formulated and designed with the aid of the Skweak [26] framework for Python, which is a toolkit to easily define, apply, and aggregate label functions. Skweak is also tightly integrated with SpaCy[27], another Python framework designed to help solve NLP problems and tokenization of texts. Finally, all Label functions were then incorporated in a script for applying everything specified above and returning the "Contracts" labeling results via IOB(inside, outside, beginning) tagging format.

Next, we present two pseudo codes, algorithms 1 and 2, for each label function, these examples were designed for the named entity "`contract_number`".

- *Regex Label Function* (algorithm 1): Searches for an occurrence of the regex expression in the act text. When found, marks the expression start and end as a "`contract_number`" named entity.

**Algorithm 1** Regex Label Function

```
1: expression ← r "Contrato (\d+)"
2: match ← regex.search(expression, actText)
3: if match then
4:     yield match.start, match.end, contract_number
5: end if
```

- *Keyword-List Label Function* (algorithm 2): Searches for an occurrence of any of the possible "starts" key-words, and if found searches for any of the "ends" key-words.

When both are found, every text between them is marked as a "`contract_number`" Named Entity.

**Algorithm 2** Keyword-List Label Function

```
starts ← ["CONTRACT", "Contract", "Contract Number"]
ends ← [",", ";", ".", "-"]
for word_count ← 0 to size(actText) do
    if actText[word_count] in starts then
        start ← actText[word_count]
        while (actText[word_count] not in ends) and (word_count < size(actText)) do
            word_count ← word_count + 1
        end while
        end ← actText[word_count]
        yield start, end, contract_number
    end if
end for
```

Next, in figure 3.4 we present a practical example referent to the weak supervision process mentioned previously, using a "Contract" text as input and showing the expected output. In the figure, we can observe how both label functions were directly applied in an unlabeled "Contract" and had their results aggregated by the HMM aggregation method.

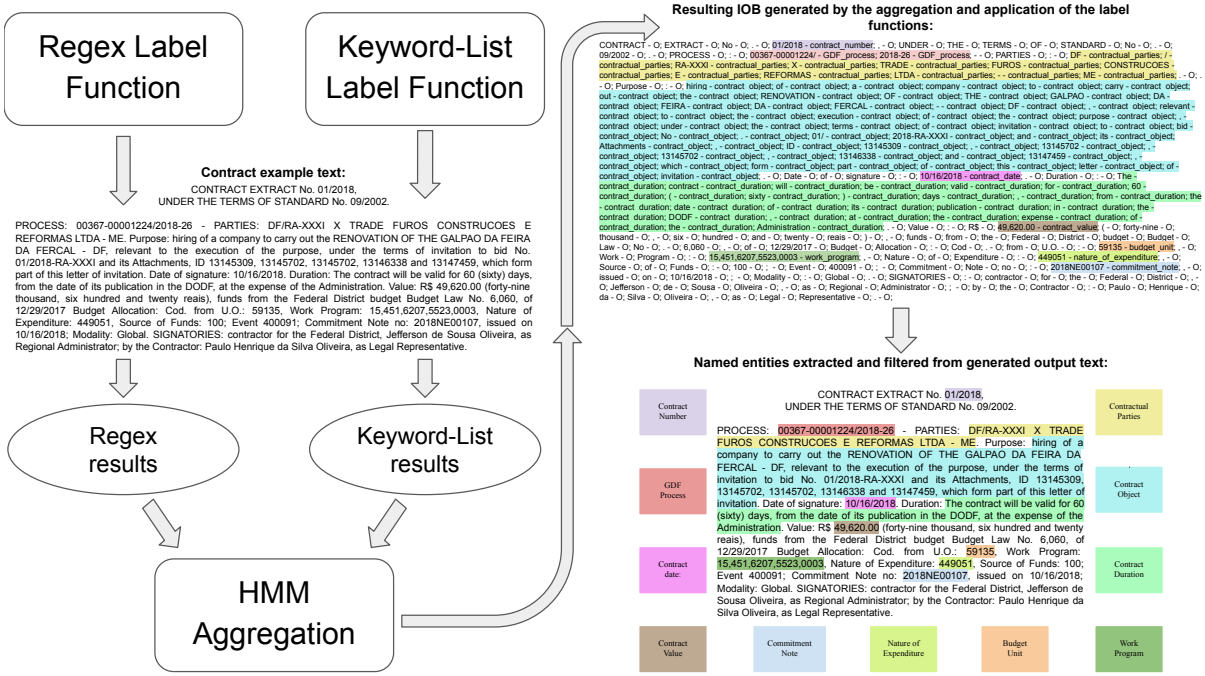


Figure 3.4: Weak Supervision labeling process example.

## 3.4 Models and Training

In relation to the NER models, we chose four pre-trained neural language models. Each model passed through a fine-tuning process [28], consisting of adding a BI-LSTM (Bidirectional Long Short-Term Memory Networks) [29] layer at its top for sequence labeling. The chosen models are the following:

- BERTimbau [30]: Pre-trained BERT model with a Brazilian Portuguese textual corpus.
- Lener-BR [1]: A fine-tuned BERTimbau model for Brazilian Portuguese legislative texts.
- RoBERTa [31]: An optimized version of the BERT model, developed with support from Facebook researchers.
- DistilBERT-PT [2]: A lighter (distilled) version of BERT, pre-trained with a Brazilian Portuguese textual corpus.

The training process itself was conducted with the aid of the ktrain framework [32], a wrapper for machine learning libraries that facilitates building and deploying neural networks. Each model was trained once in all the 783 base instances for each one of the three labeled datasets produced in the previous conducted steps, the *Human-labeled* dataset extracted from DODFCorpus I annotations, the *GPT-3-labeled* dataset extracted from GPT-3's prompt predictions, and lastly the *weak-supervision-labeled* dataset extracted from the application and aggregation of the weak supervision label functions. For training, we used a triangular learning rate policy [33], with an initial learning rate of 0.01 and a total of 10 epochs, in which we obtained the best training results.

After the usual training of the three datasets models, we conducted experimentations regarding combinations of the three labeled databases aiming to better comprehend how they could complement each other. The first experiment consisted in combining percentages (10%, 20%..., 100%) of the human-labeled base into the GPT-3 labeled base and training models for each iteration. Thus generating 10 iterations of combined models that would better display how Human data could improve training using GPT-3 labeled data, while GPT-3 would still be able to reduce some of the Human annotation effort.

The second experiment consisted in training a model with a complete combination between the GPT-3 labeled base and the weak supervision labeled base, focusing on the results both strategies would provide together. It wouldn't be effective to combine weak

[1] `https://huggingface.co/pierreguillou/bert-base-cased-pt-lenerbr`
[2] `https://huggingface.co/adalbertojunior/distilbert-portuguese-cased`

supervision annotation in percentages similar to the process conducted in the human and GPT-3 combination since the effort in this approach is present in the label functions development step not in their application into the labeling texts.

# Chapter 4

# Experimental Results

We present and discuss the experimental results considering mainly the performance overview (F1-Score) of the models trained for each of the three training bases. Table 4.1 displays the F1-Score metric for every model in every training dataset and also estimates an average score between the models by dataset. As expected, Human Labeling presented the best overall accuracy of the three databases, followed by Weak Supervision and lastly by GPT-3. This order adequately reflects the cost and effort devoted by each approach, being GPT-3 arguably the less costly technique while Human Labeling is the most expensive. It is valid to point out the specific case where *DistilBERT* Weak Supervision achieved better performance than *DistilBERT* Human Labeling.

Table 4.1: F1-Score metric and average F1-Score metric of each model in every dataset.

| | **Labeled Datasets** | | |
|---|---|---|---|
| Model | **GPT-3** | **Weak Supervision** | **Human Labeling** |
| NER-BERTimbau | 0.543 | 0.703 | 0.755 |
| NER-LenerBR | 0.554 | 0.676 | 0.761 |
| NER-RoBERTa | 0.542 | 0.674 | 0.707 |
| NER-DistilBERT-PT | 0.473 | 0.664 | 0.631 |
| Average F1-Scores | 0.528 | 0.679 | 0.713 |

By analyzing all the eleven named entities' results in Figure 4.1, we can infer that some entities represent an extremely difficult labeling task, even for the best-performing models trained using human annotations. With that in mind, and to ensure a fairer and solid analysis, we selected only the seven best performing named entities ("*contract number*", "*GDF process*", "*contract value*", "*budget unit*", "*work program*", "*nature of expenditure*", "*commitment note*") resulting from the Human Labeling training, shown in green charts of Figure 4.1, and reanalyzed the F1-Score metric.

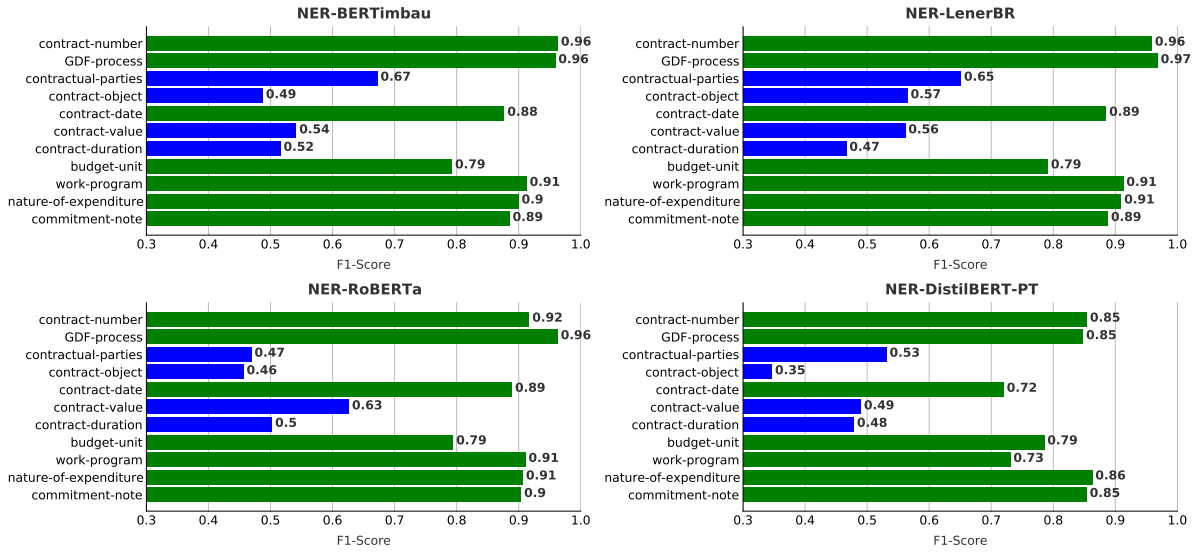


Figure 4.1: Charts representing F1-Scores of all named entities for each model trained with the human-labeled dataset. In green are the chosen seven best-performing entities.

Table 4.2 presents models F1-Score overview considering only these seven entities. Here we can observe a substantial improvement in the F1-Score values and a bigger approximation of the Weak Supervision and GPT-3 results to the Human Labeling results. Also, again *DistilBERT* Weak Supervision achieved better performance than *DistilBERT* Human Labeling.

Table 4.2: F1-Score metric considering only the seven best performing named entities.

| | **Labeled Datasets** | | |
|---|---|---|---|
| Model | **GPT-3** | **Weak Supervision** | **Human Labeling** |
| NER-BERTimbau | 0.776 | 0.887 | 0.902 |
| NER-LenerBR | 0.815 | 0.878 | 0.906 |
| NER-RoBERTa | 0.798 | 0.881 | 0.899 |
| NER-DistilBERT-PT | 0.664 | 0.847 | 0.804 |
| Average F1-Scores | 0.763 | 0.873 | 0.877 |

Considering the experiments conducted combining percentages of Human Labeling and GPT-3 Labeling, figure 4.2 presents the results on all eleven entities and figure 4.3 on the seven best-performing entities. In both figures, we can observe how each model F1-Score metric behaved with the addition of the human annotations percentages. Most of the four model charts show a similar learning curve at the starting iterations, presenting increasing improvements. After the third or fifth iteration, with more than $30\% - 50\%$ Human Labeling added, this curve usually becomes inconsistent on each iteration not always

representing gains in F1-Score. From this behavior, it is possible to infer that, at the first iterations, human data is consistently improving model performance and positively impacting GPT-3 data, while at the last iterations, it is arguable that adding more human-labeled data does not necessarily mean a definite or substantial improvement in model performance.

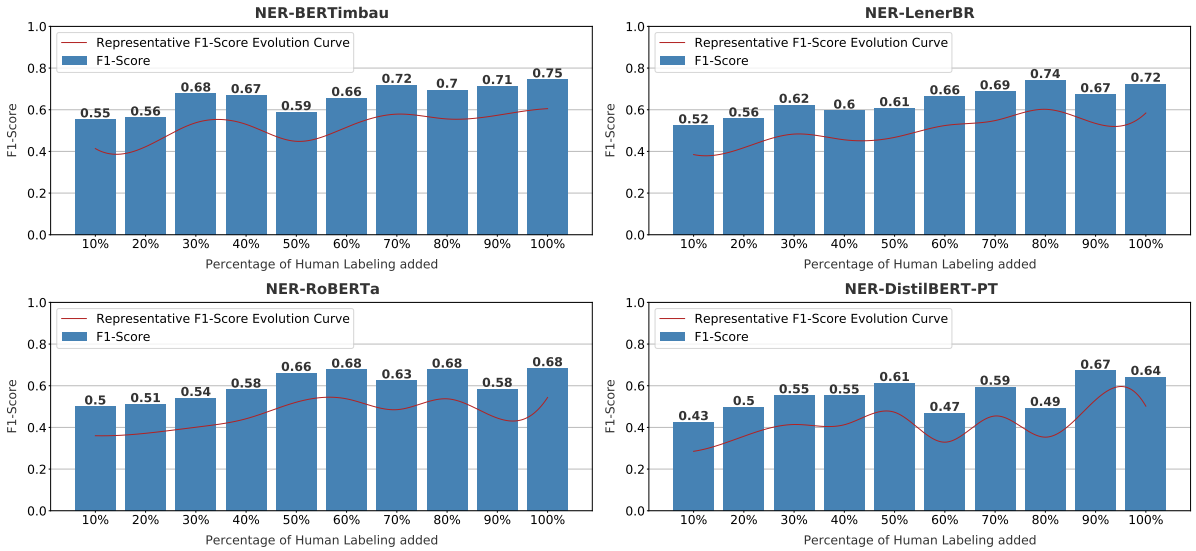


Figure 4.2: Charts representing each of the four models F1-Score over the GPT-3 and Human Labeling combining iterations.

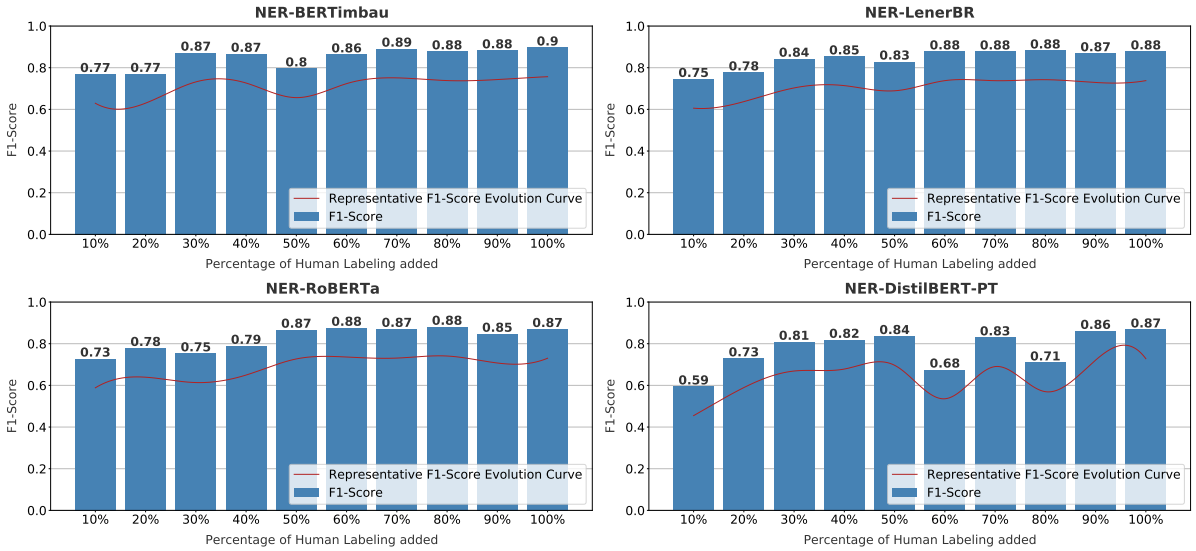


Figure 4.3: Charts representing each of the four models F1-Score over the GPT-3 and Human Labeling combining iterations considering only the seven best-performing entities.

Table 4.3 presents resulting F1-scores from GPT-3 and Weak Supervision combination. The leftmost section of the table represents the complete F1-Score overview, while the rightmost section demonstrates the overview considering only the seven best-performing named entities previously established. Results show a considerable improvement to the GPT-3 results, but also losses and very few gains when compared to Weak Supervision and Human Labeling results.

Table 4.3: F1-Score values resulting from the combination of GPT-3 and Weak Supervision datasets.

| | **GPT-3 and Weak Supervision** | |
|---|---|---|
| Model | **All Eleven Entities** | **Seven Best Entities** |
| NER-BERTimbau | 0.686 | 0.884 |
| NER-LenerBR | 0.709 | 0.888 |
| NER-RoBERTa | 0.558 | 0.773 |
| NER-DistilBERT-PT | 0.632 | 0.831 |
| Average F1-Scores | 0.646 | 0.844 |

Lastly, in table 4.4, we carried out a direct comparison between all models trained, based on a preservation score metric that focuses on their ability to preserve the performance presented by the human-trained models. The preservation score is defined as $\frac{F1\_tested\_model}{F1\_human}$, where $F1\_human$ is the performance achieved by models trained from human-labeled data and $F1\_tested\_model$ is the performance achieved by models trained with all alternative methods previously presented in this paper, that is, prompt-based data labeling (GPT-3), weak supervision (Weak-Sup), prompt-based labeling and weak supervision combination (GPT + Weak-Sup) and prompt-base labeling and human labeling combination (GPT + 30%Human).

The top part of table 4.4, presents the preservation score comparison regarding all eleven dataset named entities, while the bottom part regards only the seven best-performing named entities. High preservation score values indicate lower labeling costs while maintaining model performance. In some scenarios, the preservation score is greater than 1, thus indicating that models trained with our methods even outperformed models trained with human-annotated data.

Figure 4.4 presents the same analysis considering the preservation score, this time however, extending it to display all iterations on the combination of GPT-3 and Human Labeling percentages.

Considering table 4.4, we carried out a statistical analysis of the performance preservation scores presented considering multiple runs and the four approaches. Figure 4.5 shows the critical difference diagram for preservation score measure, computed by Friedman's

Table 4.4: Final comparison between the four methods and each of their trained models considering the preservetion score metric.

| | All Eleven Entities | | | |
|---|---|---|---|---|
| Model | **GPT-3** | **Weak-Sup** | **GPT + Weak-Sup** | **GPT + 30%Human** |
| NER-BERTimbau | 0.719 | 0.931 | 0.908 | 0.897 |
| NER-LenerBR | 0.728 | 0.888 | 0.931 | 0.818 |
| NER-RoBERTa | 0.766 | 0.953 | 0.789 | 0.765 |
| NER-DistilBERT-PT | 0.749 | 1.052 | 1.001 | 0.877 |
| Average | 0.740 | 0.956 | 0.907 | 0.839 |

| | Seven Best Performing Entities | | | |
|---|---|---|---|---|
| Model | **GPT-3** | **Weak-Sup** | **GPT + Weak-Sup** | **GPT + 30%Human** |
| NER-BERTimbau | 0.860 | 0.983 | 0.980 | 0.966 |
| NER-LenerBR | 0.899 | 0.969 | 0.980 | 0.930 |
| NER-RoBERTa | 0.887 | 0.980 | 0.859 | 0.838 |
| NER-DistilBERT-PT | 0.825 | 1.053 | 1.033 | 1.006 |
| Average | 0.867 | 0.996 | 0.963 | 0.935 |

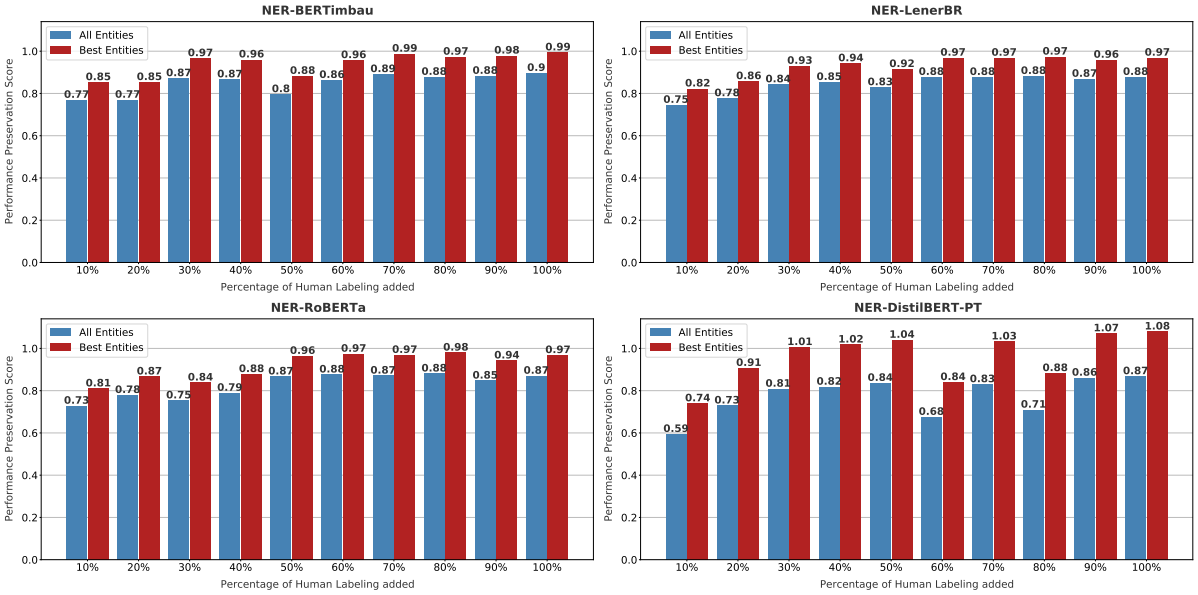


Figure 4.4: Charts representing performance preservation score of each iteration on the combination of GPT-3 and Human annotation.

test with Nemenyi's posttest with 95% confidence level, as suggested by Demšar et al. [34]. All approaches are ordered according to the average ranking of multiple runs.

The Critical Difference (CD) value obtained was 1.66, that is, there is no statistical difference in the preservation score between two methods when their difference in the average ranking is lower than this value. As shown in figure 4.5, none of the four models

presented a higher difference value in their average ranking position than CD's value, therefore we did not find a significant critical difference between the four approaches. Yet, it is essential to point out that in practical aspects, combining GPT-3 with Weak Supervision is a promising approach, as we can use GPT-3 to label the most complex entities, while the label functions can be used for entities that require simple regular expressions.

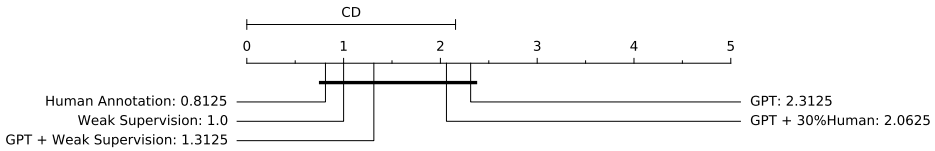


Figure 4.5: Friedman's test with Nemenyi's posttest graphical analysis.

# Chapter 5

# Conclusion

Traditional manual text labeling methods for Named Entity Recognition (NER) in legal documents are a naturally expensive task. Our paper investigates how to use prompt-based models and weak supervision as alternatives to these methods, labeling unannotated data in a cost-efficient way, and also how to train efficient deep learning models from these labeling results. We show that applying these less costly labeling strategies can still be a valid approach for training efficient Deep Neural Network Models, even with their trade-off lower performance. To address this issue, we explored three main methods and their combinations: human manual annotation, OpenAI's GPT-3 prompt-based model, and weak supervision.

Experimental results showed that human labeling still presented itself as the best approach considering model accuracy and performance, with weak supervision closely behind and GPT-3 with the worst results. This fact rightfully reflects the overall cost of each approach, with GPT-3 having the lower cost and Human Annotation having the higher cost. The combination techniques also presented considerable results that managed to further approximate human labeling performance while maintaining low cost and labeling effort.

We also established a final comparison based on a statistical analysis using Friedman's test and the preservation score metric, calculated to determine how much performance of the human-trained model was preserved by each method. Nevertheless, despite the previous findings about each model and their compared performances, we verified that there is no significant statistical difference between the four methods' preservation scores. In conclusion, all the methods presented in this article are therefore able to train, despite the variations, efficient models capable of resembling, in terms of performance and F1-Score, models trained with human data.

Limitations of this study are the absence of a precise way to estimate each approach's cost, which could bring great insights, limitations on the size of our database, and also

not considering annotator mistakes, a factor that can be very problematic in real-life scenarios. Regarding future work, we plan on expanding research to include Active Learning techniques, better integration between presented approaches possibly adding GPT-3 as one or more label functions in a weak supervision system, and designing new improved strategies for prompting GPT-3 establishing more options and metrics to select textually similar prompts for each unlabeled act, therefore, increasing GPT-3 prediction accuracy by supplying better input prompts.

# References


[1] Chowdhary, KR1442: *Natural language processing.* Fundamentals of artificial intelligence, pages 603–649, 2020. 1

[2] Torfi, Amirsina, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox: *Natural language processing advancements by deep learning: A survey.* arXiv preprint arXiv:2003.01200, 2020. 1

[3] Marrero, Mónica, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís: *Named entity recognition: fallacies, challenges and opportunities.* Computer Standards & Interfaces, 35(5):482–489, 2013. 1

[4] Giri, Rachayita, Yosha Porwal, Vaibhavi Shukla, Palak Chadha, and Rishabh Kaushal: *Approaches for information retrieval in legal documents.* In *2017 Tenth International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE, 2017. 2

[5] Sakhaee, Neda and Mark C Wilson: *Information extraction framework to build legislation network.* Artificial Intelligence and Law, 29(1):35–58, 2021. 2

[6] Nasar, Zara, Syed Waqar Jaffry, and Muhammad Kamran Malik: *Named entity recognition and relation extraction: State-of-the-art.* ACM Computing Surveys (CSUR), 54(1):1–39, 2021. 2

[7] Zhang, Shanshan, Lihong He, Eduard Dragut, and Slobodan Vucetic: *How to invest my time: Lessons from human-in-the-loop entity extraction.* In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2305–2313, 2019. 2

[8] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei: *Language models are few-shot learners.* CoRR, abs/2005.14165, 2020. `https://arxiv.org/abs/2005.14165`. 2, 9

[9] Dale, Robert: *Gpt-3: What's it good for?* Natural Language Engineering, 27(1):113–118, 2021. 2

[10] Zamani, Hamed and W. Bruce Croft: *On the theory of weak supervision for information retrieval*. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '18, page 147–154, New York, NY, USA, 2018. Association for Computing Machinery, ISBN 9781450356565. https://doi.org/10.1145/3234944.3234968. 2

[11] Zhou, Zhi Hua: *A brief introduction to weakly supervised learning*. National science review, 5(1):44–53, 2018. 3

[12] Ratner, Alexander J, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré: *Data programming: Creating large training sets, quickly*. Advances in neural information processing systems, 29, 2016. 3

[13] Bach, Stephen H., Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Chris Ré, and Rob Malkin: *Snorkel drybell: A case study in deploying weak supervision at industrial scale*. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 362–375, New York, NY, USA, 2019. Association for Computing Machinery, ISBN 9781450356435. https://doi.org/10.1145/3299869.3314036. 3

[14] Dai, Hongliang, Yangqiu Song, and Haixun Wang: *Ultra-fine entity typing with weak supervision from a masked language model*. CoRR, abs/2106.04098, 2021. https://arxiv.org/abs/2106.04098. 3

[15] Dozier, Christopher, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali: *Named entity recognition and resolution in legal text*. In *Semantic Processing of Legal Texts*, pages 27–43. Springer, 2010. 4

[16] Vardhan, Harsh, Nitish Surana, and BK Tripathy: *Named-entity recognition for legal documents*. In *International conference on advanced machine learning technologies and applications*, pages 469–479. Springer, 2021. 4

[17] Araujo, Pedro Henrique Luz de, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo: *Lener-br: a dataset for named entity recognition in brazilian legal text*. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer, 2018. 4

[18] Wang, Shuohang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng: *Want to reduce labeling cost? GPT-3 can help*. CoRR, abs/2108.13487, 2021. https://arxiv.org/abs/2108.13487. 4

[19] Meyer, Selina, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada: *Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai*. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA, 2022. Association for Computing Machinery, ISBN 9781450397391. https://doi.org/10.1145/3543829.3544529. 4

[20] Floridi, Luciano and Massimo Chiriatti: *Gpt-3: Its nature, scope, limits, and consequences*. Minds and Machines, 30(4):681–694, 2020. 5

[21] Ratner, Alexander, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré: *Snorkel: Rapid training data creation with weak supervision.* The VLDB Journal, 29(2):709–730, 2020. 5

[22] Karamanolakis, Giannis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah: *Self-training with weak supervision.* CoRR, abs/2104.05514, 2021. https://arxiv.org/abs/2104.05514. 5

[23] Lison, Pierre, Aliaksandr Hubin, Jeremy Barnes, and Samia Touileb: *Named entity recognition without labelled data: A weak supervision approach.* CoRR, abs/2004.14723, 2020. https://arxiv.org/abs/2004.14723. 5

[24] Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig: *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.* ACM Comput. Surv., 55(9), jan 2023, ISSN 0360-0300. https://doi.org/10.1145/3560815. 9

[25] Eddy, Sean R: *What is a hidden markov model?* Nature biotechnology, 22(10):1315–1316, 2004. 11

[26] Lison, Pierre, Jeremy Barnes, and Aliaksandr Hubin: *skweak: Weak supervision made easy for nlp.* arXiv preprint arXiv:2104.09683, 2021. 11

[27] Vasiliev, Yuli: *Natural Language Processing with Python and SpaCy: A Practical Introduction.* No Starch Press, 2020. 11

[28] Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang: *How to fine-tune bert for text classification?* In *China national conference on Chinese computational linguistics*, pages 194–206, Cham, 2019. Springer, Springer International Publishing. 13

[29] Graves, Alex and Jürgen Schmidhuber: *Framewise phoneme classification with bidirectional lstm and other neural network architectures.* Neural Networks, 18(5):602–610, 2005, ISSN 0893-6080. https://www.sciencedirect.com/science/article/pii/S0893608005001206, IJCNN 2005. 13

[30] Souza, Fábio, Rodrigo Nogueira, and Roberto Lotufo: *Bertimbau: pretrained bert models for brazilian portuguese.* In *Brazilian conference on intelligent systems*, pages 403–417. Springer, 2020. 13

[31] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov: *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019. 13

[32] Maiya, Arun S.: *ktrain: A low-code library for augmented machine learning.* arXiv preprint arXiv:2004.10703, 2020. 13

[33] Smith, Leslie N.: *Cyclical learning rates for training neural networks*, 2015. https://arxiv.org/abs/1506.01186. 13

[34] Demšar, Janez: *Statistical comparisons of classifiers over multiple data sets.* The Journal of Machine learning research, 7:1–30, 2006. 19

# Appendix A

# Submitted Article

This monography is a paper submitted to the Springer journal *Artificial Intelligence and Law*.

- Title: Combining prompt-based language models and weak supervision for named entity recognition from legal documents
  - Authors: Vitor Oliveira, Gabriel Nogueira, Thiago Faleiros e Ricardo Marcacini