International Journal of Combinatorial Optimization Problems and Informatics, 16(4), Sep-Dec 2025, 473-485. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i4.575

www.editada.org

Fine-Tuning Large Language Models and Machine Learning for Extracting Entities and Relationships in Spanish Medical Texts

Nidia K. Serafin-Rojas[1], *José A. Reyes-Ortiz*[1,*], *Maricela Bravo*[1] *and Josué Padilla-Cuevas*[1]
[1] Systems Department, Autonomous Metropolitan University, Azcapotzalco, Mexico City 02200, Mexico.
al2221800042@azc.uam.mx, jaro@azc.uam.mx, mcbc@azc.uam.mx, jpc@azc.uam.mx
[*] *Corresponding author*

Abstract. In this paper, we delve into the techniques used to extract information from medical texts written in Spanish. Our study focuses on fine-tuning Large Language Models by training them on Spanish medical texts and adjusting various hyperparameters. We also explore traditional machine learning algorithms like support vector machines, decision trees, and nearest neighbors. Our analysis aims to evaluate the tool's ability to identify entities and relationships between them. Our results show that the support vector machine algorithm outperformed Large Language Models in entity identification, achieving a 78.72% F_1 score compared to 68.04%. However, Large Language Models demonstrated superior performance in relationship identification, achieving a 57.15% F_1 score compared to 35.5% for machine learning algorithms.
Keywords: Natural Language Processing; Entities and Relationships Extraction; Fine-tuning Large Language Models; Machine Learning

Article Info
Received November 29, 2024
Accepted December 4, 2024

1 Introduction

The Health Sciences field generates enormous amounts of information through unstructured clinical documents, such as medical reports, diagnoses, and clinical trials. Natural Language Processing (NLP) has proven invaluable in extracting meaningful insights and relationships from this vast amount of data.

By automating the process of extracting information, NLP significantly improves the efficiency of medical data interpretation. It leads to innovation in critical areas such as clinical research, treatment development, and patient care. Therefore, it is essential to apply natural language processing techniques to unlock the potential of unstructured medical information and advance the understanding and treatment of diseases.

This study aims to compare Large Language Models (LLMs) (Abutridy, 2023) and traditional machine learning approaches (Aggarwal, 2018) for extracting information from medical texts. The process consists of collecting a set of Spanish medical texts, applying both traditional machine learning algorithms and fine-tuned Large Language Models, and evaluating their performance. The results are then analyzed to determine which approach is most effective for information extraction.

Traditional algorithms used in this study include K-nearest neighbors (KNN), Support Vector Machines (SVM), Decision Trees, and Perceptrons. A deep learning technique, Large Language Models (LLMs), will be adapted to the collected corpus.

This document is structured as follows: Section 2 reviews previous work related to traditional machine learning and deep learning; Section 3 details the dataset of medical texts used; Section 4 presents the conventional machine learning approaches employed in this study and their associated text processing; Section 5 addresses the experiments conducted

with fine-tuning large language models; Section 6 compares traditional machine learning algorithms and large language models; and finally, the conclusions are presented.

2 Work Related

This chapter provides a comprehensive review of the most relevant approaches, techniques, and methodologies used in information extraction. We analyze key contributions from traditional machine learning models to the latest deep learning architectures, highlighting their strengths, limitations, and practical applications.

A major challenge in extracting information from medical texts in Spanish is the scarcity of expert-labeled datasets (Dellanzo et al., 2022). This study underscores the need for research that demonstrates effective solutions using Spanish-language corpora. To address this gap, a corpus of 513 annotated articles was created, with the primary objective of performing named entity recognition and relationship extraction in health reports related to outbreaks in Latin America. A combination of Recurrent Neural Networks and Conditional Random Fields was employed, and the results were evaluated by category. For entity recognition, F_1 scores ranged from 10% to 77%, while relationship extraction achieved scores between 27% and 100%.

Building structured datasets is a crucial step in improving information extraction models, particularly when dealing with unstructured medical texts (Mendoza-Urbano et al., 2023). In a related effort, this study developed a natural language processing module for oncological pathology reports, requiring the construction of a dataset with 140 annotated documents. The system's performance was assessed against human-analyzed results using metrics such as accuracy, precision, recall, and F_1 score, yielding promising outcomes.

In Villena et al., (2021), an approach was proposed that employed support vector machines, random forests, logistic regression, and perceptrons to identify patients based on their medical history. This study achieved F_1 scores exceeding 90% with texts in Spanish.While much of the research in this field has been conducted in dominant languages such as English, the work by (Kaur & Khattar, 2021) distinguishes itself by focusing on Punjabi and the processing of unstructured text. The researchers implemented algorithms such as Hidden Markov Models, Maximum Entropy Principle, and Conditional Random Fields, achieving impressive F_1 scores of 77.61%, 83.65%, and 93.21%, respectively.

A separate study by Kaplar et al. (2022), researchers evaluated the performance of conditional random fields, multilingual transformers (BERT Multilingual and XLM RoBERTa), long short-term memory (LSTM) recurrent neural networks, and their ensembles for clinical entity recognition in Serbian. The study highlighted the language processing challenges in Serbian and showed promising results with the employed algorithms.

A recent study on information extraction within the medical domain (Wei et al., 2020) compared a Support Vector Machine with a deep learning approach called BiLSTM-CRF. The study concluded that the BiLSTM-CRF method, which relies on an artificial neural network, exhibited promising results. This approach was further tested by (Fang et al., 2022), which demonstrated that combining BiLSTM-CRF with transformer representations and Conditional Random Fields (BERT-BiLSTM-CRF) proved effective for named entity recognition in Chinese texts and extracting POS features.

Zhang et al. (2019) conducted a study analyzing 500 psychiatric notes in English from the Taiwan Medical Center to compare Conditional Random Fields algorithms with BLSTM-CRF neural networks. The latter was found to have better results in terms of F_1 score.

Al-Smadi et al. (2019) presented a study on sentiment analysis through various information extraction techniques. They use a Support Vector Machine for training to identify aspect categories, extract target expressions of opinion by applying natural language processing tasks, and extract morphological, syntactic, and semantic features. The study applied machine learning algorithms such as Support Vector Machines, Naive Bayes, Decision Trees, and Nearest Neighbor, using hotel reviews in Arabic as the data source.

Medical knowledge graphs are crucial in advancing healthcare applications by enabling structured representation and extraction of biomedical information. In this context, the BioIE system proposed by Wu et al. (2021) represents a significant advancement. This hybrid neural network integrates an enhanced Graph Convolutional Network (GCN) with multi-head attention mechanisms to extract relationships from biomedical texts and unstructured medical reports. The model was evaluated on two key biomedical relationship extraction tasks: chemical-disease relationships (CDR) and chemical-protein interactions (CPI), demonstrating superior performance to existing baseline models. Notably, the study leverages a corpus of cancer pathology reports collected from hospitals, reinforcing its relevance to oncology research.

Similarly, Jouffroy et al. (2021) emphasize the effectiveness of recurrent neural networks in extracting relationships between diseases and medications, including dosage, frequency, administration route, and conditions of use. A distinctive aspect of this study is the use of French-language medical texts, with the proposed model achieving a remarkable F_1 score of 89.9%, highlighting its robustness across different linguistic contexts.

Beyond the medical domain, information extraction techniques have also been applied to other fields. For instance, Kumar et al. (2022) explored their potential in agriculture by employing a Long Short-Term Memory (LSTM) neural network to extract concepts and relationships such as climate-crop interactions. The study reports performance metrics including precision, sensitivity, specificity, and F_1 score, underscoring the broad applicability of deep learning-based information extraction across diverse disciplines.

A related effort by Sánchez-Graillet et al. (2022) focuses on clinical concept extraction, utilizing a BERT-based large language model. The model was trained on medical summaries related to glaucoma and type 2 diabetes mellitus (T2DM), achieving an F_1 score of 76% for glaucoma and 77% for diabetes in an exact match evaluation. The study further emphasizes the role of ontologies in enhancing clinical concept extraction.

Finally, the studies by Yang et al. (2021) and Monteagudo-García et al. (2021) contribute to the eHealth-KD Challenge 2021 by implementing BERT-CRF, BiLSTM-CRF, and BiLSTM-based models for Task A. The reported F_1 scores were 17.3% and 60.7%, respectively. A comparative analysis in a results table indicates that the PUCRJ-PUCPR-UFMG team obtained the highest F_1 score (70.6%) in entity recognition. In comparison, the IXA team achieved the best performance in relationship identification with an F_1 score of 43%.

3 Dataset

This study focuses on analysing medical texts, necessitating the identification of datasets that experts have annotated. To achieve this, we utilize the text set provided by the eHealth-KD 2021 challenge (Piad-Morfis et al., 2021), which includes sentences sourced from various platforms such as MedlinePlus, Wikinews, and the CORD-19 corpus, all related to health topics. This diversity ensures a wide range of formats and structures.

Since we emphasize Spanish, we have adapted the original text set by removing the English sentences. The corpus consists of two files: one containing sentences with diverse structures (illustrated in Fig. 1), and the other containing annotations (shown in Fig. 2). In the annotation file, entities and relationships are identified using tags T and R, respectively, followed by a sequential number that indicates their start and endpoints. Additionally, this file specifies the class to which each identified entity or relationship belongs.

Original text	Text translated into English
El sistema vascular es la red de vasos sanguíneos del cuerpo.	The vascular system is the network of blood vessels in the body.

Fig. 1. Example of the text file with its English translation from (Piad-Morfis et al., 2021)

Original text		Text translated into English	
T1	Concept 3 10;11 19 sistema vascular	T1	Concept 3 10;11 19 vascular system
T2	Predicate 26 29 red	T2	Predicate 26 29 network
T3	Concept 33 38;39 49 vasos sanguíneos	T3	Concept 33 38;39 49 blood vessels
T4	Concept 54 60 cuerpo	T4	Concept 54 60 body
R0	arg Arg1:T2 Arg2:T3	R0	arg Arg1:T2 Arg2:T3
R1	is-a Arg1:T1 Arg2:T2	R1	is-a Arg1:T1 Arg2:T2
R2	part-of Arg1:T1 Arg2:T4	R2	part-of Arg1:T1 Arg2:T4

Fig. 2. Example of the annotations file with its English translation from (Piad-Morfis et al., 2021)

3.1 Entities

Entities (Real Academia Española, n.d.) are elements of value or importance in something. Generally, in the context of information extraction, they are specific and distinctive elements identified and extracted from a text or corpus of data. These include people, organizations, places, medical terms, and other relevant information about the analyzed domain. This study focuses on extracting general entities, which are classified into four categories: concepts, actions, predicates, and references.

Table 1 shows the frequency of appearances by entity type in the content of the analyzed sentences.

Table 1. Entity Type Frequency

Entity Type	Frequency
Action	512
Concept	1388
Predicate	190
Reference	55
Total	2145

3.2 Relationships

The dataset includes information about relations, defined as "connection, correspondence of something with another thing" (Real Academia Española., n.d.). In this case, among the entities. In this work, the search for 12 types of relations is employed.

Table 2. Entity Relationship Frequency

Relation Type	Frequency
target	154
in-context	98
subject	86
is-a	75
domain	40
causes	38
in-place	38
arg	31
in-time	22
has-property	14

part-of	10
entails	9
Total	615

The relations can be classified as general type, which means they are only relationships between two entities, such as is-a, has-property, part-of, causes, and entails. For example: *has-property*: indicates an entity has a specific property or characteristic.

Contextual relationships allow refining an entity, such as in-time, in-place, and in-context. For example: "exposición / exposition" in-time "verano / summer".

Action role type relations indicate entities roles in action, such as subject and target. For example: "[el] asma afecta [. . .]} / [the] asthma affects [. . .] ".

Moreover, predicate role type relations indicate entities' roles in a predicate, such as domain and arg. An example of this type of relation would be as in "mayores [de] 60 años / older [than] 60 years".

4 Text processing

Traditional machine learning algorithms require applying Natural Language Processing techniques to convert text into a numerical matrix, which acts as input for the algorithms. In this study, Python is the primary tool used to extract relevant sentences from the text dataset.

4.1 Feature extraction

Feature extraction produces a numerical matrix in which morphological features are derived from part-of-speech (POS) tagging, while semantic features are obtained through Named Entity Recognition (NER) tagging. This process was executed utilizing the Python library Spacy.

The segmented text was extracted from the dataset. For entity identification, the input includes a sentence containing the entity of interest, accompanied by its generated context, which comprises a window of five words to the left and five words to the right of the entity. This meticulous method allows us to capture the essential context surrounding each entity, underscoring the importance of our research.

We also implemented a similarly practical approach for relationship identification. The sentence containing the first term related to the relationship is extracted and extended until the second term is located. This robust method ensures that entity relationships are accurately captured, highlighting the practicality of our findings.

Table 3 and Table 4 provide examples of extracted entities and relationships from the relevant text, helping to understand this extraction process.

Table 3. Extraction of Relevant Text for Entities (Piad-Morfis et al., 2021)

Entity	Relevant Text	Entity Type
sistema vascular / vascular system	*El sistema vascular es la red de vasos sanguíneos del cuerpo.* / The vascular system is the network of blood vessels in the body.	Concept
red / network	*El sistema vascular es la red de vasos sanguíneos del cuerpo.* / The vascular system is the network of blood vessels in the body.	Predicate

elevado / elevated	*La columna de cenizas se ha elevado por más de 18.000 pies*. / The ash column has risen to over 18,000 feet.	Action
Estos / these	*Estos síntomas pueden conducir a pérdida.* / These symptoms may lead to loss.	Reference

Table 4. Extraction of Relevant Text for Relations (Piad-Morfis et al., 2021)

Relation	Entity 1	Entity 2	Relevant Text
arg	*red* / network	*vasos sanguíneos* / blood vessels	*red de vasos sanguíneos* / Network of blood vessels.
is-a	*sistema vascular* / vascular system	*red* / network	*sistema vascular es la re* The vascular system is the network
part-of	*sistema vascular* / vascular system	*cuerpo* / body	*sistema vascular es la red de vasos sanguíneos del cuerpo* / The vascular system is the network of blood vessels in the body.

4.2 Machine Learning algorithms

The algorithms selected for this work, which focus on identifying entities and relationships, were chosen for their diverse characteristics (Aggarwal, 2018). They are implemented in Python using the scikit-learn library, as outlined below. Several widely used machine learning algorithms are typically employed in data classification and regression tasks.

4.2.1 K-Nearest Neighbors

The nearest neighbor algorithm, K-Nearest Neighbors or k-NN, is a supervised machine learning technique for classification and regression tasks. In classification, the algorithm utilizes a training dataset with established labels to assign a class label to a new instance. This classification process involves examining the labels of the k nearest instances in the training set.

It is important to note that the choice of k can significantly influence the algorithm's accuracy. In this instance, a value of k=7 has been utilized, meaning that the labels of the seven closest instances are considered for classification. However, this value can be adjusted based on the problem's specific requirements and the data's characteristics. Overall, the k-NN algorithm is a versatile tool capable of addressing various classification and regression challenges.

4.2.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that excels at classification tasks by identifying the optimal hyperplane that effectively separates data points into distinct classes. It employs a regularization parameter set to 1.0 and utilizes a Gaussian kernel with a degree of 3. This methodology is implemented through the scikit-learn SVC library.

4.2.3 Decision Trees

In a decision tree, each internal node represents a specific feature or attribute, while the branches extending from that node indicate the possible values or categories of that feature. Leaf nodes symbolize the final classes or output values.

The *DecisionTreeClassifier* from scikit-learn is utilized to construct the tree, which involves partitioning the dataset based on attributes to achieve the best class separation and maximize uncertainty reduction.

4.2.4 Perceptron

The Perceptron is an artificial neuron model that combines several weighted inputs using an activation function. The developed code uses the default values of the *scikit-learn* Perceptron implementation, including alpha=0.0001, max_iter=1000, eta0=1.0, tol=1e-3, and shuffle=True.

5 Entities and relations extraction using fine-tuning large language models

Large language models have emerged as a transformative technology in Natural Language Processing (NLP), significantly enhancing our interactions with information. These models, built upon deep neural networks and trained on extensive datasets, demonstrate exceptional performance across various tasks, including text classification, generation, and translation.

Among these advanced models, BERT (Bidirectional Encoder Representations from Transformers), developed by Google, is particularly noteworthy.

This study focuses on BERT-based models, such as RoBERTa: A Robustly Optimized BERT Pretraining Approach and the BETO Spanish model, detailed below.

5.1 Base BERT Model

BERT's unique ability to capture bidirectional context within a sentence means it can understand the meaning of a word and its surrounding words. Thanks to its attention mechanism, BERT has proven effective in various natural language processing (NLP) applications. It is exceptionally skilled at identifying and categorizing entities such as names, organizations, and locations in text data, even when embedded in complex text structures.

5.2 RoBERTa Model

RoBERTa is a language model developed by Facebook AI, grounded in the Transformer architecture (Liu, 2019). It represents an enhanced and optimized iteration of the BERT model and is widely recognized for its exceptional performance across various natural language processing (NLP) tasks.

A key feature of *RoBERTa* is its emphasis on pre-training the model using extensive datasets. It leverages a substantial volume of unlabeled text to train the model on a self-regression task, where the model predicts the subsequent words in a text sequence based on the preceding context. Additionally, essential hyperparameters, such as batch size and epochs, can be fine-tuned. These advancements have significantly boosted the model's performance in various NLP applications, including text classification, information extraction, question answering, and text generation.

5.3 BETO Model

BETO, an acronym for "Bidirectional Encoder Representations from Transformers for Spanish" (Canete, 2023), is a pre-trained language model specifically designed for the Spanish language. It is built on the Transformer architecture, similar to well-known models like *BERT* and *RoBERTa*. *BETO* operates bidirectionally and employs a pretraining strategy that enables it to learn contextualized word representations from a large corpus of unlabeled Spanish text.

BETO's pretraining process uses a diverse and extensive dataset encompassing various domains and writing styles in Spanish. This approach allows the model to effectively capture a range of grammatical structures, vocabulary, and semantics unique to the Spanish language.

5.4 Fine-tuning large language model

Fine-tuning large language models (LLMs) is a crucial technique in natural language processing. These models are initially trained on extensive datasets of general-purpose text, and fine-tuning allows them to adapt to specific tasks. In this study, training is conducted using a collection of medical texts in Spanish, accompanied by hyperparameter tuning.

Additionally, the presence of pre-trained models significantly accelerates the fine-tuning process. This research focuses on refining the models *PlanTL-GOB-ES/roberta-base-biomedical-clinical-es* and *plncmm/beto-clinical-wl-es,* the latter being a modified version of the *dccuchile/bert-base-spanish-wwm-uncased* model.

Their architecture and size do not solely determine the effectiveness; fine-tuning plays a pivotal role in tailoring them for specific tasks, such as information extraction. This process involves adjusting the model's weights and hyperparameters according to a particular dataset relevant to the task, thereby enhancing performance.

This study investigates manual hyperparameter tuning for large language models aimed at information extraction. Two pre-trained models, utilizing Spanish texts, are employed. Through fine-tuning, we can unlock the full potential of these models and effectively extract valuable information from vast amounts of textual data.

Below is a detailed manual calibration of the *plncmm/beto-clinical-wl-es* and *PlanTL-GOB-ES/roberta-base-biomedical-clinical-es* models, incorporating experiments with three adjustments outlined in Table 5. These models were also re-trained using the texts from this research as part of the classification task.

Table 5. Hyperparameters obtained through exploratory search, manual tuning, and trial and error.

Hyperparameters	Configuration 1	Configuration 2	Configuration 3
Batch size	8	8	8
Learning rate	0.00005	0.00005	0.00005
Weight decay	0	0.01	0.01
Optimizer	adamw_hf	adafactor	adagrad
Epoch	3	1	3

In addition to these configurations, the model was re-trained with the text corpus used in this study.

6 Experimental Results

This section delineates the comprehensive evaluation of models developed for targeted tasks. The findings elucidate the performance metrics of the respective algorithms and models, with particular emphasis on their efficacy when applied to novel, unseen data.

6.1 Metrics

As discussed in Section 3, the classes within this text corpus are imbalanced. Consequently, employing a metric that provides a balanced evaluation of the model's performance is crucial. In this study, we choose the F_1 score as our evaluation metric. The F_1 score, representing the harmonic mean of precision and recall, considers false positives and false negatives.

The F_1 score is calculated using Equation (3), while precision and recall are derived from the computations in Equation (1) and Equation (2). By thoroughly assessing the model's performance, the F_1 score is a fitting metric for our analysis.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (1)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (2)$$

$$F_1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

6.2 Results of Machine Learning Approaches

In this study, we conducted experiments using Natural Language Processing (NLP) techniques to analyze texts and extract both morphological and semantic features. We then applied traditional machine learning algorithms, including Nearest Neighbor, Support Vector Machines, Decision Trees, and Perceptrons, to classify the data. We used the F_1 score metric to evaluate our models' performance.

6.2.1 Results of Machine Learning Approaches for Entity Identification

We extracted features of entities and their contexts from the text corpus and converted this information into a numerical matrix representing label frequencies. This matrix is used as input for the machine learning algorithms. The models were evaluated based on their F_1, and the results are presented in Table 6.

Table 6. Results with traditional algorithms for entity identification.

Algorithm	POS-tagged Entity (%)	POS-tagged Context (%)	NER-tagged Context (%)	POS Entity + POS Context + NER Context (%)
KNN	76.97	52.99	52.66	72.13
SVM	77.22	48.91	53.04	78.72
Decision Trees	76.43	55.42	55.42	69.51
Perceptron	71.14	32.12	52.20	70.49

6.2.2 Results of Machine Learning Approaches for Relations Identification

This section presents the results obtained by applying specific techniques and algorithms to identify relationships between entities in the analyzed text corpus using the F_1 score metric, as shown in Table 7.

Table 7. Results with traditional algorithms for entity identification.

Algorithm	POS-tagged Context (%)	NER-tagged Context (%)	POS Entity + POS Context + NER Context (%)
KNN	28.81	10.21	34.29
SVM	26.56	17.45	35.50
Decision Trees	29.12	18.38	30.96
Perceptron	14.20	5.40	13.23

6.3 Results of fine-tuning large language models

This section presents the findings derived from the application and evaluation of the large language model (LLM) for analyzing entities and relations.

6.3.1 Entities

The experimental outcomes pertaining to entity identification utilizing large language models are delineated below, exhibiting performance metrics ranging from 68.04% to 49.49%. Presented in Table 8 are the results corresponding to the models configured as discussed in Section 5, which refer to the specifications detailed in Table 5. The designated hyperparameters have been meticulously employed in this study.

Table **8.** Results with large language models, pretrained and a combination of configurations from Section 5, for entity identification.

Experiment	Model	Score F_1(%)
1	Configuration 1 + beto-clinical-wl-es	68.04
2	Configuration 2 + beto-clinical-wl-es	67.34
3	Configuration 3 + beto-clinical-wl-es	66.73
4	Configuration 1 + roberta-base-biomedical-clinical-es	49.49
5	Configuration 2 + roberta-base-biomedical-clinical-es	49.49
6	Configuration 3 + roberta-base-biomedical-clinical-es	49.49

6.3.1 Relationships

The same hyperparameter configuration was implemented for the relationships, resulting in experiments with an F_1 score ranging from 20.39% to 57.15%, as detailed in Table 9.

Table 9. Results with pre-trained and fine-tuned large language models for relation identification.

Experiment	Model	Score F_1(%)
1	Configuration 1 + beto-clinical-wl-es	54.42
2	Configuration 2 + beto-clinical-wl-es	57.15
3	Configuration 3 + beto-clinical-wl-es	56.24
4	Configuration 1 + roberta-base-biomedical-clinical-es	20.39
5	Configuration 2 + roberta-base-biomedical-clinical-es	25.71
6	Configuration 3 + roberta-base-biomedical-clinical-es	25.30

6.4 Analysis and Discussion of Results

This section compares the most significant results, concentrating on evaluating the F_1 Score. These outcomes were derived from the application of machine learning algorithms and are contrasted with those achieved using large language models for entity and relation identification.

6.4.1 Entity Identification

We begin by evaluating the performance of machine learning (ML) and large language model (LLM) algorithms in entity identification, explicitly focusing on concepts, entities, actions, and references. Our analysis utilizes the F_1 score metric for each approach to assess their effectiveness in capturing entities within the medical domain, as detailed in Table 10.

Table 10. Comparative results between the best ML and LLM models for entity identification.

Model	Score F_1 (%)
SVM + POS Entity + POS Context + NER Context	78.72
Configuration 1 + beto-clinical-wl-es	68.04

The results of this analysis indicate that the Support Vector Machine (SVM) algorithm notably excelled, achieving an F_1 score of 78.72% for this task. This outcome underscores the efficacy of traditional machine learning methods in entity identification, reinforcing their ongoing relevance in extracting information from clinical texts. The SVM's effectiveness in this scenario can be attributed to the relatively small size of the text corpus.

Upon reviewing the findings, we observe that the performance gap between the two methods is approximately 10%. This margin presents an opportunity for improvement in both approaches. For example, obtaining more relevant features could enhance the F_1 score for the SVM method. Expanding the dataset size and exploring hyperparameter optimization for the LLM approach could also improve performance.

6.4.2 Relations Identification

In the subsequent phase of our research, we focus on identifying relationships between entities within medical texts. This study aims to evaluate the efficacy of machine learning (ML) and large language models (LLMs) in detecting and classifying these relationships.

Table 11. Comparative results between the best ML and LLM models for entity identification.

Model	Score F_1 (%)
SVM + POS Entity + POS Context + NER Context	35.50
Configuration 2 + beto-clinical-wl-es	57.15

The findings indicate that the LLMs achieved superior performance in this domain. While the text corpus utilized was relatively small, it encompassed a diverse array of categories, precisely 12 distinct, notably imbalanced types. As a result, support vector machine (SVM) models may necessitate a greater volume of training examples across each category to enhance their predictive accuracy.

Conversely, traditional machine learning algorithms outperformed the large language models in entity analysis. The application of the SVM algorithm yielded more favorable results, mainly attributable to the size of the text set employed in the study. Furthermore, it is important to note that only four distinct categories of entities were identified.

In contrast, evaluating relationships yielded contrasting results, with the large language model demonstrating exceptional capability. It can be attributed to the fact that the number of relationship categories significantly exceeded that of entity categories, with a ratio of approximately three to one. Additionally, since traditional machine learning algorithms lack pre-training and the instances per category required balancing, achieving satisfactory performance levels was a time-consuming process.

7 Conclusions and Future Work

This paper explores the application of traditional machine learning algorithms and large language models (LLMs) to extract information from Spanish medical texts. The findings indicate that traditional machine learning algorithms, particularly Support Vector Machines (SVMs), excelled in entity identification, achieving an F_1 score of 78.72%, which surpasses the performance of LLMs. However, LLMs demonstrated a more remarkable ability to identify relationships within the texts, securing an F_1 score of 57.15%, compared to the 35.5% achieved by traditional methods.

These results highlight the nuanced effectiveness of different approaches depending on the specific aspect of information extraction being targeted. While traditional methods are robust for entity recognition, LLMs' advanced contextual understanding capabilities provide significant advantages in discerning complex relationships in unstructured data.

The research emphasizes the critical role of machine learning in enhancing the interpretability of medical documentation. Future improvements in algorithmic approaches could be attained through more refined feature extraction and hyperparameter tuning. Comparing the two methodologies offers valuable insights into their strengths and limitations, paving the way for future studies to optimize information extraction technologies for medical applications.

For future work, there is an intention to calibrate the hyperparameters of machine learning algorithms and identify a broader range of features within medical literature texts. Additionally, experiments with various large language models and hyperparameters, as well as with other text sets, are planned.

References

Aggarwal, C. C. (2018). Machine learning for text. Springer International Publishing. https://doi.org/10.1007/978-3-319-73531-3.

Abutridy, J. A. (2023). Grandes modelos de lenguaje: Conceptos, técnicas y aplicaciones. Alfaomega – Marcombo

Dellanzo, A., Cotik, V., Lozano Barriga, D. Y., Mollapaza Apaza, J. J., Palomino, D., Schiaffino, F., ... & Ochoa-Luna, J. (2022). Digital surveillance in Latin American diseases outbreaks: information extraction from a novel Spanish corpus. BMC bioinformatics, 23(1), 558.

Mendoza-Urbano, D. M., Garcia, J. F., Moreno, J. S., & Juan Carlos, J. (2023). Extracción automatizada de información en español de texto libre de informes de patología oncológica. *Colombia Médica*, 54(1), 1-12.

Villena, F., Pérez, J., Lagos, R., & Dunstan, J. (2021). Supporting the classification of patients in public hospitals in Chile by designing, deploying and validating a system based on natural language processing. *BMC medical informatics and decision making*, 21, 1-11.

Kaur, A., & Khattar, S. (2021, September). A systematic exposition of Punjabi Named Entity Recognition using different Machine Learning models. In 2021 *Third International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1625-1628). IEEE.

Kaplar, A., Stošović, M., Kaplar, A., Brković, V., Naumović, R., & Kovačević, A. (2022). Evaluation of clinical named entity recognition methods for Serbian electronic health records. *International Journal of Medical Informatics,* 164, 104805.

Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., ... & Xu, H. (2020). A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association,* 27(1), 13-21.

Fang, A., Hu, J., Zhao, W., Feng, M., Fu, J., Feng, S., ... & Chen, X. (2022). Extracting clinical named entity for pituitary adenomas from Chinese electronic medical records. *BMC medical informatics and decision makin*g, 22(1), 72.

Zhang, Y. C., Chung, W. C., Li, K. H., Huang, C. C., Dai, H. J., Wu, C. S., ... & Liang, T. Y. (2019, July). Depressive Symptoms and Functional Impairments Extraction From Electronic Health Records. In 2019 *International Conference on Machine Learning and Cybernetics (ICMLC)* (pp. 1-6). IEEE.

Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., & Qawasmeh, O. (2019). Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management*, 56(2), 308-319.

Wu, J., Zhang, R., Gong, T., Liu, Y., Wang, C., & Li, C. (2021, December). Bioie: Biomedical information extraction with multi-head attention enhanced graph convolutional network. In 2021 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2080-2087). IEEE.

Jouffroy, J., Feldman, S. F., Lerner, I., Rance, B., Burgun, A., & Neuraz, A. (2021). Hybrid deep learning for medication-related information extraction from clinical texts in French: MedExt algorithm development study. *JMIR medical informatics*, 9(3), e17934.

Kumar, S., Sastry, H. G., Marriboyina, V., Alshazly, H., Idris, S. A., Verma, M., & Kaur, M. (2022). Semantic Information Extraction from Multi-Corpora Using Deep Learning. *Computers, Materials & Continua,* 70(3).

Sanchez-Graillet, O., Witte, C., Grimm, F., & Cimiano, P. (2022). An annotated corpus of clinical trial publications supporting schema-based relational information extraction. *Journal of Biomedical Semantics*, 13(1), 14.
Yang, M. (2021). Yunnan-1 at eHealth-KD Challenge 2021: Deep-Learning Methods for Entity Recognition in Medical Text. In *IberLEF@ SEPLN* (pp. 725-730).
Monteagudo-Garcıa, L., Marrero-Santos, A., Fernández-Arias, M. S., & Canizares-Dıaz, H. (2021). Uh-mmm at ehealth-kd challenge 2021. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021).*
Piad-Morffis, A., Gutiérrez, Y., Cañizares-Diaz, H., Estevez-Velarde, S., Muñoz, R., Montoyo, A., & Almeida-Cruz, Y. (2020). *Overview of the ehealth knowledge discovery challenge at iberlef 2020.*
Real Academia Española. (n.d.). Entidad | Diccionario de la lengua española. Edición del Tricentenario. Recuperado el 22 de marzo de 2024, de https://dle.rae.es/entidad
Real Academia Española. (n.d.). Relación | Diccionario de la lengua española. Edición del Tricentenario. Recuperado el 22 de marzo de 2024, de https://dle.rae.es/relación#JsL0ZPA
Hugging Face. (n.d.). PlanTL-GOB-ES/roberta-base-biomedical-clinical-es. The AI community building the future. Recuperado el 22 de marzo de 2024, de https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es
Hugging Face. (n.d.). plncmm/beto-clinical-wl-es. The AI community building the future. Recuperado el 22 de marzo de 2024, de https://huggingface.co/plncmm/beto-clinical-wl-es
Hugging Face. (n.d.). Trainer. The AI community building the future. Recuperado el 22 de marzo de 2024, de https://huggingface.co/docs/transformers/v4.19.2/en/main_classes/trainer#transformers.TrainingArguments
Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.
Liu, Y. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364.
Cañete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2023). *Spanish pre-trained bert model and* evaluation *data. arXiv preprint arXiv*:2308.02976.