Article

Weakly-Supervised Multilingual Medical NER For Symptom Extraction For Low-Resource Languages

Rigon Sallauka [*] , Umut Arioz , Matej Rojc , Izidor Mlakar

Posted Date: 16 April 2025

doi: 10.20944/preprints202504.1356.v1

Keywords: low-resource languages; machine translation; medical entity extraction; NER; NLP; patient-reported outcomes; weakly-supervised learning doi:10.20944/preprints202504.1356.v1

Article

Weakly-Supervised Multilingual Medical NER For Symptom Extraction For Low-Resource Languages

Rigon Sallauka *, Umut Arioz, Matej Rojc and Izidor Mlakar

HUMADEX Group, Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroška cesta 46, 2000 Maribor, Slovenia; rigon.sallauka@um.si; umut.arioz@um.si; matej.rojc@um.si; izidor.mlakar@um.si

* Correspondence: rigon.sallauka@um.si

Abstract: Patient-reported health data, especially Patient-Reported Outcomes Measures, are vital for improving clinical care but are often limited by memory bias, cognitive load, and inflexible questionnaires. Patients prefer conversational symptom reporting, highlighting the need for robust methods in symptom extraction and conversational intelligence. This study presents a weakly-supervised pipeline for training and evaluating medical Named Entity Recognition (NER) models across eight languages, with a focus on low-resource settings. A merged English medical corpus, annotated using the Stanza i2b2 model, was translated into German, Greek, Spanish, Italian, Portuguese, Polish, and Slovenian, preserving entity annotations (PROBLEM, TEST, TREATMENT). Data augmentation addressed class imbalance, and fine-tuned BERT-based models consistently outperformed baselines. The English model achieved the highest F1 score (80.07%), followed by German (78.70%), Spanish (77.61%), Portuguese (77.21%), Slovenian (75.72%), Italian (75.60%), Polish (75.56%), and Greek (69.10%). Compared to existing baselines, our models demonstrated notable performance gains, particularly in English, Spanish, and Italian. This research underscores the feasibility and effectiveness of weakly supervised, multilingual approaches for medical entity extraction, contributing to improved information access in clinical narratives—especially in under-resourced languages.

Keywords: low-resource languages; machine translation; medical entity extraction; NER; NLP; patient-reported outcomes; weakly-supervised learning

1. Introduction

With the growing emphasis on patient-centered care, patient-reported outcomes (PROMs) have emerged as a critical tool for understanding and addressing patients' health needs, especially in the management of chronic and complex conditions [1,2]. PROMs allow the systematic collection of patients' health experiences, symptoms, and quality of life, often using structured instruments such as questionnaires [3,4]. While effective in many cases, traditional PROMs face notable limitations. These include their reliance on structured, predefined questions that may not fully capture patients' experiences, a lack of adaptability for underrepresented and resource-poor languages, and the significant burden they can place on patients to complete standardized forms [5,6]. Moreover, current PROMs often fail to accommodate natural language expression, which could offer richer and more nuanced insights into patients' symptoms and health concerns [7]. This issue is particularly pronounced in diverse linguistic contexts, where a lack of tools for resource-poor languages limits inclusivity and equity in data collection and analysis [8,9].

In light of the limitations of traditional PROMs, there is a growing need for approaches that can capture patient experiences in a more natural and inclusive manner, accommodating diverse linguistic contexts and minimizing patient burden [10]. To address these challenges, advancements in Natural Language Processing (NLP), such as Named Entity Recognition (NER), offer promising tools for extracting patient-reported information directly from unstructured text, enabling a richer and more equitable understanding of health concerns across languages. NER is a fundamental NLP task that involves extracting entities from text and classifying them into predefined categories, such as locations, people, or organizations [11]. In the medical domain, clinical NER models have been developed to identify specific information like diseases, drugs, dosages, and frequencies, with recent advancements leveraging deep learning algorithms to enhance their performance [11]. While progress has been made in medical NER for resource-abundant languages, there remains a notable research gap concerning resource-poor languages [12]. Addressing this gap is important to ensure equitable healthcare advancements across various language groups. Moreover, beyond structuring medical notes, there is a need to extract symptoms or problems as expressed by patients in their own words, facilitating a more patient-centered approach to healthcare.

This paper describes a weekly-supervised pipeline for training and evaluating medical NER models across eight languages: English, German, Greek, Spanish, Italian, Polish, Portuguese, and Slovenian. By fine-tuning BERT based models, we enabled the extraction of medical entities like symptoms, treatments, and tests from medical and everyday texts. The models will be used in a multicentric feasibility study of SMILE project, as part of digital health interventions to support (mental) well-being of children across multiple sites in Germany, Italy, UK, Spain, Cyprus, Poland, and Slovenia. This research focuses on underrepresented languages and patient-expressed symptoms, as it seeks to enhance the accessibility and quality of healthcare information extraction.

Clinical Named Entity Recognition (NER) models have made significant progress in structuring and summarizing clinical notes, predominantly in English. However, existing models face several challenges, including monolingual focus, inconsistent performance reporting, issues with non-clinical semantics in reporting, and limited adaptability to low-resource languages [13–15]. The model in silpakanneganti/bert-medical-ner (https://huggingface.co/silpakanneganti/bert-medical-ner) on HuggingFace (https://huggingface.co/) platform extracts the entities, such as: age, sex, clinical event, nonbiological location, duration, severity, biological structure, sign symptom, and more, achieving a self-reported F1 score of 67.5%. Similarly, samrawal/bert-base-uncased_clinical-ner (https://huggingface.co/samrawal/bert-base-uncased_clinical-ner) is another NER model, which extracts problem, treatment, and test entities from medical texts but it does not report any performance evaluation, leaving its effectiveness unclear. Other models, such as the one proposed in [16] and Stanza's clinical i2b2 model [17,18], report F1 scores of 81.2% and 88.1%, respectively, yet their focus remains restricted to English medical texts. Domain-specific models like ugaray96/biobert_ncbi_disease_ner (https://huggingface.co/ugaray96/biobert_ncbi_disease_ner) and Kaelan/en_ner_bc5cdr_md (https://huggingface.co/Kaelan/en_ner_bc5cdr_md) are tailored for tasks like disease or chemical recognition, reporting F1 scores of 85.7% for chemicals and diseases. Moreover, they are well-tuned to medical semantics but not to the way individuals express or describe their symptoms. Efforts in non-English languages include models like GERNERMED [19] for German, which extracts entities such as drug dosage and frequency (F1: 81.5%), MedPsyNIT [20] for Italian, which extracts symptoms and comorbidities (F1: 89.5%), and the model in [21] for Spanish, which recognizes anatomical, chemical, and pathological entities (F1: 86.4%). The model for Portuguese achieves a high F1 score of 92.6% [22] for extracting entities like disorders, medical procedures, and pharmacologic substances. Despite these successes, multilingual models for medical NER remain scarce, and no models currently exist for Greek, Polish, or Slovenian.

In summary, while existing clinical NER models demonstrate strong performance in specific settings, they are constrained by their focus on a single language or a narrow range of languages, their reliance on extensive annotated datasets that are expensive and time-intensive to create, and their limited ability to adapt across different types of texts, such as formal medical documents and informal everyday language.

Our contributions are as follows:

- we introduce a flexible, multilingual, and adaptable pipeline for medical NER

- our pipeline enables effective extraction of medical entities (problem, test and treatment) from a given body of text
- our pipeline is effective across diverse languages, with a focus on low-resource languages our pipeline is designed to handle both formal and informal textual contexts.

2. Materials and Methods

2.1. Overall Methodology

In this paper, we propose a weakly supervised multilingual Named Entity Recognition (NER) pipeline for symptom extraction from medical texts and patient reported data. The pipeline consists of several key stages: data preparation, annotation, translation for multilingual adaptation, data augmentation and model fine-tuning. Each step contributes to building a robust NER model capable of identifying symptom-related entities across eight languages (English, German, Greek, Italian, Spanish, Polish, Portuguese, and Slovenian) in unstructured medical texts. The flow of the pipeline is displayed in the image below, providing an overview of the processes involved in creating the final model. Figure 1 shows the overview of the pipeline flow.

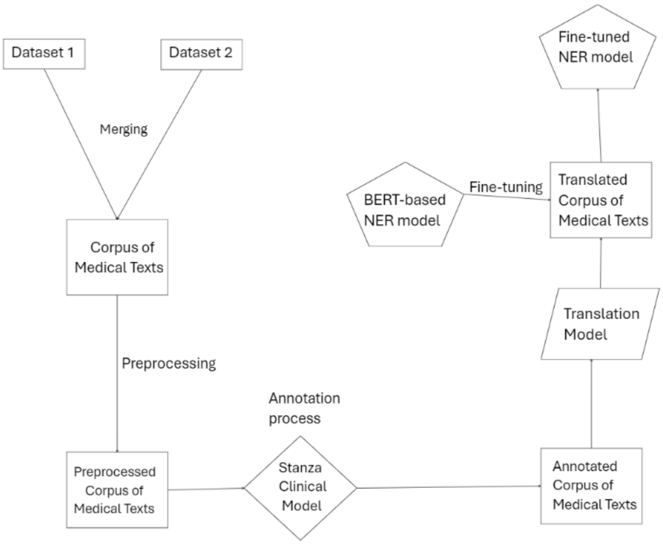

Figure 1. Overview of the pipeline flow.

2.2. Data Sources and Corpus Creation

The initial stage involved combining multiple datasets of medical texts to form a comprehensive corpus of medical texts. This corpus was created by merging two key datasets Kabatubare/autotrain-data-1w6s-u4vt-i7yo and s200862/medical_qa_meds. The dataset Kabatubare/autotrain-data-1w6s-u4vt-i7yo consists of clinical notes, discharge summaries, and other unstructured medical records from various healthcare providers, providing a rich source of terminology related to patient symptoms, diagnoses, and treatments. The dataset in s200862/medical_qa_meds, on the other hand, contains a diverse collection of question-and-answer pairs focused on medical topics, which includes descriptions of symptoms, diagnoses, and treatment options. By merging these datasets, we ensured a diverse and comprehensive representation of medical terminology and symptom descriptions in English, capturing different styles of clinical language, patient narratives, and medical abbreviations. The total number of examples of both datasets is 29,379 entries of text data.

Details of dataset Kabatubare/autotrain-data-1w6s-u4vt-i7yo:

```
dataset_info:
  features:
    - name: autotrain_text
      dtype: string
  splits:
    - name: train
      num_bytes: 19,109,937
      num_examples: 23,437
```

Details of dataset s200862/medical_qa_meds:

```
dataset_info:
  features:
    - name: text
      dtype: string
  splits:
    - name: train
      num_bytes: 1,638,400
      num_examples: 5,942
```

2.3. Preprocessing

The data underwent a thorough preprocessing phase. Initially, it was cleaned to remove unnecessary text while retaining relevant information. Since our dataset consisted of question-answer pairs between a user and an assistant, we identified certain textual elements that could be removed. In the dataset in Kabatubare/autotrain-data-1w6s-u4vt-i7yo, we removed strings such as "Human:", "Assistant:", "\n", "\t", and replaced hyphens ("-") between words with a single space. In the dataset in s200862/medical_qa_meds, we eliminated tags like "[INST]", "[/INST]", "<s>", "</s>", along with newline ("\n") and tab ("\t") characters. Next, all punctuation was removed from the text. Finally, the data was converted to lowercase for consistency. As stated above, the rows of the datasets are medical texts, made of multiple sentences. We split all the texts as to have one sentence at each row, thus the number of examples rose from around 29,000 to 261,936 sentences.

2.4. Annotation Process

To annotate the corpus of medical texts, we utilized Stanza [18], a Python NLP package designed for linguistic analysis across multiple human languages. Stanza offers a suite of efficient tools that perform a variety of linguistic processing tasks, including sentence segmentation, word tokenization, part-of-speech tagging, and entity recognition. Starting from raw text, Stanza structures the data, identifying syntactic and semantic elements to support further analysis.

In particular, Stanza provides an i2b2 model specialized for NER tasks in the medical domain. This model is designed to extract key medical entities: PROBLEM (symptom), TEST, and TREATMENT. Trained on the i2b2-2010 corpus [30], which includes manually annotated clinical notes from institutions, such as Partners Healthcare, Beth Israel Deaconess Medical Center, and the University of Pittsburgh Medical Center, the i2b2 model is highly accurate and well-suited for medical text analysis. With an impressive F1-score of 88.13, the i2b2 model effectively identifies and classifies medical entities, making it a robust tool for NER tasks in clinical settings. Table 1 presents i2b2 model from Stanza.

Table 1. Stanza model's description [18].

Model	Dataset	Entity Types	Number of (train/dev/test)	Tokens	F1-score
i2b2 clinical model	i2b2-2010 corpus	Problem, Test, Treatment	106,597/44,672/269,954		88.13

2.5. Translation and Multilingual Adaptation

To enable multilingual capabilities in our NER pipeline, we employed machine translation models to convert the annotated English medical texts into several target languages, including German, Greek, Spanish, Italian, Polish, Portuguese, and Slovenian. These translations allowed us to develop a dataset that supports multilingual symptom extraction, enhancing the model's utility across diverse linguistic contexts.

For the translations from English to German, Greek, Spanish, Italian, Polish, and Portuguese, we utilized models developed by the Language Technology Research Group at the University of Helsinki [23]. These models, licensed under CC-BY-4.0, are optimized for direct translations from English to each target language and are available on the Hugging Face platform [17]. Each model was specifically selected for its accuracy and reliability in medical contexts, given the specific terminology present in clinical texts:

- German - Helsinki-NLP/opus-mt-en-de (https://huggingface.co/Helsinki-NLP/opus-mt-en-de)
- Greek - Helsinki-NLP/opus-mt-en-el (https://huggingface.co/Helsinki-NLP/opus-mt-en-el)
- Spanish - Helsinki-NLP/opus-mt-tc-big-en-es (https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-es)
- Italian - Helsinki-NLP/opus-mt-tc-big-en-it (https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-it)
- Polish - gsarti/opus-mt-tc-en-pl (https://huggingface.co/gsarti/opus-mt-tc-en-pl)
- Portuguese - Helsinki-NLP/opus-mt-tc-big-en-pt (https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-pt)
- Slovenian - facebook/mbart-large-50-many-to-many-mmt model (https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt)

For English to Slovenian translations, we used the facebook/mbart-large-50-many-to-many-mmt model [24], a fine-tuned version of the mBART-large-50 model capable of translating between any pair of 50 languages. We chose this model due to its superior performance in handling low-resource languages like Slovenian. This model enables direct translation by setting the target language ID as the first generated token in the output sequence. This flexibility and multilingual adaptability were crucial in creating a reliable Slovenian translation.

Helsinki-NLP does not support the Slovenian language, that is why the mBART model was chosen for the translations in this language, while Greek is not supported in the mBART model and that is why the Helsinki-NLP was chosen for this language. Other languages are supported by both models. We chose the Helsinki-NLP model for the translation in these languages because of the higher BLEU score that it has compared to mBART model. Table 2 shows the BLEU scores for German, Spanish, Italian, Polish, and Portuguese of both models.

2.6. Word Alignment

To ensure accurate word alignment in translated texts, we used the model aneuraz/awesome-align-with-co(https://huggingface.co/aneuraz/awesome-align-with-co), which leverages contextual word embeddings from models like BERT to map words between languages effectively [25]. The alignment process involved several key steps. Contextual word embeddings were generated for each word in a sentence to capture the word's meaning based on its context within the sentence. To calculate alignment scores, two techniques were employed. First, probability thresholding involved generating a similarity matrix using the dot product of word embeddings, which was then converted into probabilities through a function like softmax. This approach identified high-probability word pairs as aligned. Second, optimal transport treated alignment as a transportation problem, minimizing the "cost" of transferring probability mass between words in different languages and producing a matrix of probable alignments. Then, alignments were calculated in both directions—source-to-target and target-to-source—with the final alignment derived from the intersection of these bidirectional results. For words split into subwords by the model, alignment was ensured through any matching subwords, guaranteeing complete word coverage in the alignment process.

2.7. Data Augmentation

Before splitting the dataset into train/validate/test sets for the training process, we performed data augmentation to enhance the diversity and robustness of the training data. Data augmentation was used because the dataset contained a large number of zero tags (O), that is the entities that do not belong to any category (PROBLEM, TEST, and TREATMENT), and this leads to class imbalance, which impacts the model’s performance. The augmentation process involved two main strategies. Sentence reordering was applied, where words within each sentence were rearranged to create new variations of the same sentence structure. This method increased the variability of the dataset, enabling the model to generalize better to different sentence formations. Additionally, entity extraction was performed by identifying all words within each sentence that were annotated with non-"O" labels (i.e., labeled as PROBLEM, TEST, or TREATMENT). These extracted words were used to generate new sentences, which were then added back into the dataset. This ensured that the model would encounter more examples of key medical entities during training.

Table 2. BLEU scores for SPECIFIC languages of Facebook MBART and Helsinki NLP machine translation models.

	Facebook MBART (ML-FT N to 1)	**Helsinki NLP**
De	41.5	**47.5**
El	/	**56.4**
Es	28.6	**54.9**
It	43.9	**53.9**
Pl	32.9	**47.5**
Pt	49.3	**50.4**
Sl	**33.9**	/

Notes: Multilingual finetuning (ML-FT); Many to one (N to 1); German (De); Greek (El); Spanish (Es); Italian (It); Polish (Pl);.

After the data augmentation, each dataset consisted of approximately 440,000 rows.

2.8. Data Splitting

Following augmentation, the dataset was split into three distinct sets. The training set, comprising 80% of the dataset, the validation set, accounting for 10% of the dataset, and the remaining 10% was designated as the test set.

2.9. Model Configuration

For our symptom extraction pipeline, we fine-tuned the BERT base model (cased), a widely-used transformer model pretrained on English text using a masked language modeling (MLM) objective [26]. This model, developed by Google, is case-sensitive, differentiating between lowercase and uppercase words (e.g., "english" vs. "English"). BERT was pretrained on large English corpora, including BookCorpus and English Wikipedia, using two main objectives: MLM, where 15% of the words in each sentence are masked and predicted, and Next Sentence Prediction (NSP), which trains the model to predict sentence order [26]. The model configuration includes a vocabulary size of 30,000, a maximum token length of 512, and an Adam optimizer with learning rate warmup and linear decay [26]. BERT achieves high performance on a range of NLP tasks, including an average score of 79.6 on the GLUE benchmark, and is well-suited for token classification tasks, such as named entity recognition, when fine-tuned on specific labeled datasets [26]. This makes it an ideal base for developing a specialized model for medical named entity recognition in our pipeline.

The model was trained to classify entities into specific categories: PROBLEM, TEST, and TREATMENT, using the IOB tagging scheme (Inside-Outside-Beginning) [27], a widely-used standard in NER, where additional tags (e.g., B-, I-, E-, S-) are used to capture entity boundaries. Key training parameters were set as follows. The architecture utilized was BERT-base-cased, a pretrained transformer model with 12 layers, 768 hidden units, and a vocabulary of 30,000 tokens.

The original architecture of the model remained unchanged. For training parameters, the model was trained for 200 epochs with a batch size of 64. An AdamW optimizer was used, featuring a learning rate of 3e-5 and a weight decay of 0.01 to prevent overfitting. The sequence length was capped at 128 tokens to ensure efficient processing. To address class imbalance, a focal loss function [28] was applied, emphasizing harder-to-classify samples.

Table 3 displays more details on the parameters used for model training.

These configurations were selected to maximize the model's ability to accurately extract medical entities from medical texts while maintaining efficiency and generalizability.

Table 3. Hyperparameters for model training.

	BERT-BASE-CASED
max sequence length	128
batch size	64
learning rate	3e-05
warmup steps	66,540
checkpoint every	2000
weight decay	0.01
max number of train epochs	200
layers	12
hidden states	768
attention heads	12
vocab size	28,996
train time (hours)	12-23
loss	focal loss
number GPUs	1
GPU type	NVIDIA A100-PCIE-40GB

2.10. Evaluation

The fine-tuned BERT-based NER model was evaluated using standard metrics: Precision, Recall, and F1-score (Table 4), which provide insights into the model's ability to accurately identify and classify medical entities. Early stopping was employed to prevent overfitting, halting training if validation loss did not improve for 30 evaluation steps.

Performance was monitored on a validation set every 2000 steps, with the best-performing model checkpoint saved based on the lowest validation loss. These metrics, along with confusion matrices, offered a detailed view of the model's performance across entity classes. The model configuration, training hyperparameters, and evaluation metrics were consistent across all eight training processes for each language. The best models of each language and the respective datasets are uploaded to the Hugging Face platform in the HUMADEX page.

Table 4. Evaluation metrics.

Metric	Formula
Precision (P)	P = TP / (TP + FP)
Recall (R)	R = TP / (TP + FN)
F1-Score (F1)	F1 = 2 * (P * R) / (P + R)

Notes: TP – True Positives; FP – False Positives; FN – False Negatives.

3. Results

This section presents the evaluation results of the eight fine-tuned multilingual NER models, each trained to extract medical entities (PROBLEM, TEST, TREATMENT) in different languages. The performance of each model is analyzed using standard metrics, including Precision, Recall, and F1-score.

Table 5 provides a comparative overview of the performance metrics for the eight language-specific NER models. The English model demonstrated the highest overall performance, with a Precision of 80.85%, Recall of 79.30%, and F1 score of 80.07%, alongside the lowest Evaluation Loss at 0.24. This suggests that the English model is both accurate and consistent in identifying entities, likely due to the abundance of high-quality English training data. The German and Spanish models also showed strong results, with F1 scores of 78.70% and 77.61%, respectively, and evaluation losses below 0.35, indicating reliable performance in these languages. Portuguese followed closely with an F1 score of 77.21%, showing its capability in medical NER despite slight variations in Recall and Precision.

In contrast, the Greek model exhibited the lowest F1 score at 69.10% and a higher evaluation loss of 0.41, suggesting greater challenges in accurately recognizing entities in Greek. The models for Italian, Polish, and Slovenian fell within a mid-range performance level, with F1 scores between 75.56% and 75.72% and evaluation losses ranging from 0.34 to 0.40. These results indicate that while the models perform reasonably well, there may be language-specific nuances or variations in translation quality that affect performance.

The Translation BLEU score column, represents the BLEU score for translations from English to each respective language, measuring the quality of machine translation used to generate training data in different languages.

Table 5. Overview of the models' performance.

Language	Precision	Recall	F1 score	Loss	BLEU score
En	80.85%	79.30%	**80.07%**	0.24	N/A

De	78.94%	78.46%	**78.70%**	0.30	47.5
El	70.69%	67.58%	**69.10%**	0.41	56.4
Es	77.14%	78.08%	**77.61%**	0.33	54.9
It	75.91%	75.28%	**75.60%**	0.34	53.9
Pl	75.52%	75.60%	**75.56%**	0.40	47.5
Pt	77.25%	77.16%	**77.21%**	0.34	50.4
Sl	75.78%	75.66%	**75.72%**	0.37	33.9

Notes: English (En); German (De); Greek (El); Spanish (Es); Italian (It); Polish (Pl); Portuguese (Pt); Slovenian (Sl).

3.1. Models' Comparison with Existing Models

To evaluate the effectiveness of our multilingual NER models, we compared their performance with three existing baseline models trained for English, Italian, and Spanish NER tasks. These existing models were tested on the same dataset used for our models to ensure a fair and consistent evaluation.

In our test set, our models outperformed the existing ones across all three languages, demonstrating superior Precision, Recall, and F1 scores, as shown in Table 6. For English, our model achieved an F1 score of 80.07%, surpassing the baseline model's performance of 67.69%, which is the Stanza model [18] by a significant margin. Similarly, in Italian, our model attained an F1 score of 75.60%, compared to the baseline model IVN-RIN/MedPsyNIT [20], which achieved an F1 score of 57.06%. In Spanish, our model's F1 score of 77.61% showed a clear improvement in handling complex linguistic nuances over the existing model lcampillos/roberta-es-clinical-trials-ner [21], which achieved an F1 score of 62.60%.

Table 6. Comparison of our fine-tuned models with existing models. Existing models are the state-of-the-art models found in the literature for specific models.

Language	Existing model	Fine-tuned model
English	67.69%	**80.07%**
Italian	57.06%	**75.60%**
Spanish	62.60%	**77.61%**

Notes: Fine-tuned models are BERT-based models, that we retrained using our datasets. The percentages are F1-scores.

The Italian and Spanish models use different labels from our labels. To do the comparison we have done label mapping to match the labels in Spanish model, as displayed in Table 7.

Table 7. Mapping of the labels of the Spanish model to labels we use in fine-tuned model.

Label in Spanish model	Mapped label
DISO	PROBLEM
PROC	TREATMENT

CHEM	TREATMENT

Whereas Table 8 shows the mapping of the labels of the Italian model.

Table 8. Mapping of the labels of the Italian model to labels we use in fine-tuned model.

Label in Italian	**Mapped label**
SINTOMI COGNITIVI	PROBLEM
TRATTAMENTO FARMACOLOGICO	TREATMENT
DIAGNOSI E COMORBIDITA	PROBLEM
SINTOMI NEUROPSICHIATRICI	PROBLEM
TEST	TEST

3.2. Case Studies

To illustrate the performance of our models, we present examples of sentences with the extracted entities categorized into PROBLEM, TEST, and TREATMENT. These examples showcase the practical application of our models in identifying relevant medical entities within unstructured text, emphasizing their ability to handle diverse linguistic constructs across multiple languages. Fine-tuned models are BERT-based models, that we retrained using our datasets.

English sentences

Fine-tuned model predictions

1. The patient complained of severe (B-PROBLEM) headaches (E-PROBLEM) and nausea (S-PROBEM) that had persisted for two days. To alleviate the (B-PROBLEM) symptoms (E-PROBLEM), he was prescribed paracetamol (S-TREATMENT) and advised to rest and drink plenty of fluids.
2. The patient exhibited symptoms (S-PROBLEM) of fever (S-PROBLEM), cough (S-PROBLEM), and body (B-PROBLEM) aches (E-PROBLEM). A (B-TEST) chest (I-TEST) X-ray (I-TEST) was taken to rule out pneumonia (S-PROBLEM). He was prescribed an (B-TREATMENT) antibiotic (E-TREATMENT) and advised to rest.
3. The patient complained of dizziness (S-PROBLEM), vision (B-PROBLEM) disturbances (E-PROBLEM), and numbness (B-PROBLEM) in (I-PROBLEM) her (I-PROBLEM) hands (E-PROBLEM). An (B-TEST) MRI (I-TEST) of (I-TEST) the (I-TEST) brain (E-TEST) was ordered to rule out a (B-PROBLEM) neurological (I-PROBLEM) cause (E-PROBLEM). A (B-TREATMENT) beta-blocker (I-TREATMENT) was prescribed to stabilize her (B-TEST) blood (I-TEST) pressure (E-TEST).

Existing model predictions

1. The patient complained of severe (B-PROBLEM) headaches (E-PROBLEM) and nausea (S-PROBEM) that had persisted for two days. To alleviate the (B-PROBLEM) symptoms (E-PROBLEM), he was prescribed paracetamol (S-TREATMENT) and advised to rest and drink plenty of fluids.
2. The patient exhibited symptoms (S-PROBLEM) of fever (S-PROBLEM), cough (S-PROBLEM), and body (B-PROBLEM) aches (E-PROBLEM). A (B-TEST) chest (I-TEST) X-ray (I-TEST) was taken to rule out pneumonia (S-PROBLEM). He was prescribed an (B-TREATMENT) antibiotic (E-TREATMENT) and advised to rest.
3. The patient complained of dizziness (S-PROBLEM), vision (B-PROBLEM) disturbances (E-PROBLEM), and numbness (B-PROBLEM) in (I-PROBLEM) her (I-PROBLEM) hands (E-PROBLEM). An (B-TEST) MRI (I-TEST) of (I-TEST) the (I-TEST) brain (E-TEST) was ordered to rule out a (B-PROBLEM) neurological (I-PROBLEM) cause (E-PROBLEM). A (B-TREATMENT) beta (I-

doi:10.20944/preprints202504.1356.v1

TREATMENT) – (I-TREATMENT) blocker (E-TREATMENT) was prescribed to stabilize her (B-TEST) blood (I-TEST) pressure (E-TEST).

The fine-tuned model demonstrates superior handling of “colloquial” symptom descriptions. When processing clinical terminology like "neurological cause" and "beta-blocker," both models maintain precise entity boundaries and correct classification. For non-clinical expressions, such as, "body aches" instead of "myalgia," the model successfully identifies these as PROBLEM entities, showing adaptability to patient-level language. In contrast, the existing model has issues with compound terms and informal expressions, particularly in maintaining consistent entity boundaries.

Spanish sentences

Fine-tuned model predictions

1. El paciente se quejó de fuertes (B-PROBLEM) dolores (E-PROBLEM) de cabeza y náuseas (S-PROBLEM) que habían persistido durante dos días. Para aliviar los síntomas, se le recetó paracetamol (S-TREATMENT) y se le aconsejó descansar y beber muchos líquidos.
2. El paciente presentó síntomas (S-PROBLEM) de fiebre (S-PROBLEM), tos y dolores (E-PROBLEM) corporals (E-PROBLEM). Se le realizó una (B-TEST) radiografía (E-TEST) de tórax para descartar una (B-PROBLEM) neumonía (E-PROBLEM). Se le recetó un (B-TREATMENT) antibiótico (E-TREATMENT) y se le aconsejó descansar.
3. La paciente se quejó de mareos (S-PROBLEM), alteraciones (E-PROBLEM) de la vision (B-PROBLEM) y entumecimiento (B-PROBLEM) en (I-PROBLEM) las manos (E-PROBLEM). Se ordenó una (B-TEST) resonancia (I-TEST) magnética (E-TEST) del (I-TEST) cerebro (E-TEST) para descartar una causa (E-PROBLEM) neurológica (I-PROBLEM). Se le recetó un (B-TREATMENT) betabloqueante (E-TREATMENT) para estabilizar su (B-TEST) presión (E-TEST) arterial (I-TEST).

Existing model predictions

1. El paciente se quejó de fuertes dolores (B-DISO) de (I-DISO) cabeza (I-PROBLEM) y náuseas (B-DISO) que habían persistido durante dos días. Para aliviar los síntomas (B-PROBLEM), se le recetó paracetamol (B-CHEM) y se le aconsejó descansar (B-PROC) y beber muchos líquidos.
2. El paciente presentó síntomas (B-DISO) de (I-DISO) fiebre (I-DISO), tos (I-DISO) y dolores (B-DISO) corporals (I-DISO). Se le realizó una radiografía (B-PROC) de (I-PROC) tórax (I-PROC) para descartar una neumonía (B-DISO). Se le recetó (B-PROC) un antibiótico (B-CHEM) y se le aconsejó descansar (B-PROC).
3. La paciente se quejó de mareos (B-DISO), alteraciones (B-DISO) de (I-DISO) la (I-DISO) vision (I-DISO) y entumecimiento (B-DISO) en (I-DISO) las (I-DISO) manos (I-DISO). Se ordenó una resonancia (B-PROC) magnética (I-PROC) del (I-PROC) cerebro (I-PROC) para descartar una causa neurológica. Se le recetó un betabloqueante (B-CHEM) para estabilizar (B-PROC) su presión (B-PROC) arterial (I-PROC).

English translation of the Spanish sentences:

1. The patient complained of severe headaches and nausea that had persisted for two days. To relieve the symptoms, paracetamol was prescribed, and the patient was advised to rest and drink plenty of fluids.
2. The patient presented symptoms of fever, cough, and body aches. A chest X-ray was performed to rule out pneumonia. An antibiotic was prescribed, and the patient was advised to rest.
3. The patient complained of dizziness, vision disturbances, and numbness in the hands. An MRI of the brain was ordered to rule out a neurological cause. A beta-blocker was prescribed to stabilize her blood pressure.

The distinction between clinical and colloquial language is especially evident in Spanish examples. The model proposed in this paper effectively processes both formal medical terms ("resonancia magnética" - MRI) and everyday expressions ("dolores de cabeza" - headaches) as coherent entities. The existing model shows a bias toward clinical terminology, using a rigid DISO/PROC/CHEM classification that poorly accommodates natural patient expressions. For instance, "mareos" (dizziness) is correctly identified as a PROBLEM by our model but gets an overly clinical DISO tag in the existing model.

Italian sentences

Fine-tuned model predictions

1. Il paziente ha lamentato forti (B-PROBLEM) mal (E-PROBLEM) di testa (E-PROBLEM) e nausea (S-PROBLEM) che persistevano da due giorni. Per alleviare i sintomi (E-PROBLEM), gli è stato prescritto il paracetamolo (S-TREATMENT) e gli è stato consigliato di riposare e bere molti liquidi.
2. Il paziente ha manifestato sintomi (S-PROBLEM) di febbre (S-PROBLEM), tosse (S-PROBLEM) e dolori (E-PROBLEM) muscolari. È stata eseguita una (B-TEST) radiografia (E-TEST) del torace (E-TEST) per escludere una (B-PROBLEM) polmonite (E-PROBLEM). Gli è stato prescritto un (B-TREATMENT) antibiotico (E-TREATMENT) e gli è stato consigliato di riposare.
3. La paziente ha lamentato vertiginid (S-PROBLEM), disturbi (E-PROBLEM) visivi (B-PROBLEM) e intorpidimento (B-PROBLEM) delle (I-PROBLEM) mani (E-PROBLEM). È stata ordinata una (B-TREATMENT) risonanza (I-TEST) magnetica (I-TEST) del (I-TEST) cervello per escludere una (B-PROBLEM) causa (E-PROBLEM) neurologica (I-PROBLEM). È stato prescritto un (B-TREATMENT) betabloccante (E-TREATMENT) per stabilizzare la pressione (E-TEST) sanguigna.

Existing model predictions

1. Il paziente ha lamentato forti mal di testa e nausea che persistevano da due giorni. Per alleviare i sintomi, gli è stato prescritto il paracetamolo (TRATTAMENTO FARMACOLOGICO (B)) e gli è stato consigliato di riposare e bere molti liquidi.
2. Il paziente ha manifestato sintomi di febbre, tosse e dolori muscolari. È stata eseguita una radiografia (TEST (B)) del torace per escludere una polmonite. Gli è stato prescritto un antibiotico e gli è stato consigliato di riposare.
3. La paziente ha lamentato vertigini, disturbi visivi e intorpidimento delle mani. È stata ordinata una risonanza (TEST (B)) magnetica (TEST (B)) del (TEST (I)) cervello (TEST (I)) per escludere una causa neurologica. È stato prescritto un betabloccante (TRATTAMENTO FARMACOLOGICO (B)) per stabilizzare la pressione sanguigna.

English translation of the Italian sentences:

1. The patient complained of severe headaches and nausea that had persisted for two days. To relieve the symptoms, paracetamol was prescribed, and he was advised to rest and drink plenty of fluids.
2. The patient presented symptoms of fever, cough, and muscle aches. A chest X-ray was performed to rule out pneumonia. An antibiotic was prescribed, and he was advised to rest.
3. The patient complained of dizziness, visual disturbances, and numbness in the hands. An MRI of the brain was ordered to rule out a neurological cause. A beta-blocker was prescribed to stabilize blood pressure.

Both formal and informal medical expressions in Italian reveal key differences between the models. The model proposed in this paper successfully identifies colloquial symptom descriptions like "mal di testa" (headache) as PROBLEM entities while maintaining accuracy with clinical terms like "betabloccante" (beta-blocker). The existing model, however, shows a clear preference for formal medical terminology, however, often missing informal symptom descriptions entirely.

4. Discussion

The presented weakly-supervised multilingual NER pipeline addresses key limitations identified in existing approaches, including single-language focus [20,24], reliance on large, annotated datasets, inability to handle informal patient language [5–7], and limited adaptation to resource-poor languages [12–14]. Namely, traditional medical NER models like Stanza [17] and GERNERMED [19] demonstrate effectiveness in single languages but struggle with multilingual adaptability [29]. Moreover, existing multilingual models often depend on extensive annotated datasets, making them impractical for resource-poor languages [12,13]. The proposed approach overcomes these constraints through weak supervision and efficient language-language translation pipelines, enabling robust performance across multiple languages, including underrepresented ones such as Slovenian and Polish.

A critical challenge in current medical NER models is their limited adaptability across different types of language [10,15]. Traditional medical NER systems, trained primarily on clinical documentation, often struggle to recognize symptoms when patients use colloquial expressions or metaphorical language to describe their conditions [5,8]. Namely, while models like Stanza [17] and MedPsyNIT [20] achieve high performance on formal clinical notes (F1 scores of 88.1% and 89.5% respectively), their effectiveness diminishes significantly when processing informal patient descriptions. This limitation is particularly evident when trying to exploit the medical NER concept in the context of patient-reported outcomes, where individuals describe their symptoms in natural, conversational language rather than clinical terminology [6,7]. For instance, while a clinician might document "pyrexia," patients typically describe "feeling hot" or "burning up." We demonstrate this with a series of case studies, comparing existing medical NERs in English, Italian and Spanish, where our model significantly overcomes the gap between clinical precision and patient expression, making it more suitable for processing real-world patient-reported health data across languages. Moreover, the increased performance over existing models, especially on the non-medical terminology (English: 67.69% to 80.07%, Italian: 57.06% to 75.60%, Spanish: 62.60% to 77.61%), further demonstrates the effectiveness of our approach. The results clearly show that our models (and the pipeline) can efficiently complement the traditional PROMs by extracting symptoms from natural language descriptions, supporting more patient-centered data collection. This well aligns with recent research highlighting the importance of natural language processing in healthcare [10], while extending its applicability to multiple languages.

The current hyperparameter configuration, while effective, still has room for optimization through more extensive tuning. Secondly, the quality of machine translation affects model performance, particularly for low-resource languages with less reliable translation tools. This limitation connects to broader challenges in multilingual NLP noted by Zhu et al. [14]. Thirdly, while our automated annotation process using Stanza [17] enables efficient dataset creation, it may propagate biases from the base model. A fully manual annotation process would provide more reliable training data but require prohibitive time and cost investments. Moreover, the reliance on pretrained BERT architecture means inherent biases in the original training data may impact performance, especially for rare medical terms or linguistic nuances in low-resource languages. Finally, class imbalance remains one of the main challenges and the less frequent entity types like TEST and TREATMENT may be underrepresented, particularly in languages with fewer examples. This echoes similar challenges noted in recent clinical NLP research [15]. While current augmentation helps balance the O tags versus medical entity tags, more advanced techniques could better balance between PROBLEM, TEST, and TREATMENT categories themselves. Moreover, sophisticated augmentation could introduce controlled semantic variations in how symptoms are described, helping models better handle the gap between clinical and patient language [5]. Furthermore, advanced augmentation techniques could generate language-specific variations that account for cultural and linguistic differences in symptom description [13], [14], particularly important for the SMILE project's multicenter implementation.

5. Conclusions

This research demonstrates the potential of fine-tuned BERT-based NER models for multilingual medical text analysis, achieving strong and consistent performance across eight languages, including underrepresented ones such as Slovenian, Polish, and Greek. By outperforming existing models in English, Italian, and Spanish, our approach highlights the importance of fine-tuning, custom annotations, and multilingual training processes in addressing the challenges of Named Entity Recognition across diverse linguistic contexts. By facilitating the extraction of detailed patient-reported symptoms and medical entities from Patient-Reported Outcomes (PROs), this research addresses the limitations of traditional structured approaches, offering a more inclusive and natural language-driven method for healthcare data collection. Despite challenges such as reliance on pretrained models, potential biases in annotation processes, translation inaccuracies, and class imbalances, the models developed in this research demonstrate strong adaptability and practicality for real-world healthcare applications. Future research should focus on refining translation pipelines, enhancing the quality of training data, and conducting extensive hyperparameter tuning to improve performance further. These efforts could expand the applicability of the pipeline to additional languages and tasks, contributing to more equitable and efficient multilingual healthcare information extraction worldwide. Finaly future work should consider developing more sophisticated data augmentation techniques. While current augmentation focuses on sentence reordering and entity extraction, more advanced techniques could synthesize natural patient language patterns, further improving the model's ability to handle informal symptom descriptions. Moreover, this could significantly reduce reliance on translation and improve ross-lingual robustness.

Author Contributions: conceptualization: IM, RS; data curation: RS; methodology: IM, RS; software: IM, RS, MR, UA; validation: IM, RS; formal analysis: IM, RS, UA, MR, investigation: IM, RS writing—original draft preparation: RS, IM; writing—review and editing: IM, RS, UA, MR; visualization: RS; supervision: IM; funding acquisition: IM ; All authors have read and agreed to the published version of the manuscript.

Funding: This paper has received funding from European Union Horizon Europe Research and Innovation Program project SMILE (grant number 101080923) and from Marie Skłodowska-Curie Doctoral Networks Actions (HORIZON-MSCA-2021-DN-01-01) project BosomShield (grant number 101073222). The content of this paper does not reflect the official opinion of the European Union or any other institution. Responsibility for the information and views expressed herein lies entirely with the authors. The funder had no impact on the design or execution of this research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets and models are available in the Zenodo platform, accessible in this link: https://zenodo.org/records/13918323 (https://doi.org/10.5281/zenodo.13918323).

Acknowledgments: During the preparation of this work the authors used ChatGPT 4o and Gemini to support the writing process as language editing tools. After using these tools/services, the authors carefully reviewed and edited the content and take full responsibility for the content of the published article.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Abbreviations

The following abbreviations are used in this manuscript:

NER	Named Entity Recognition
PROM	Patient-Reported Outcome Measure
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
MLM	Masked language modeling
NSP	Next Sentence Prediction

References

1. S. W. Salmond and M. Echevarria, “Healthcare Transformation and Changing Roles for Nursing,” Orthopaedic Nursing, vol. 36, no. 1, p. 12, Feb. 2017. https://doi.org/10.1097/NOR.0000000000000308.
2. G. Anderson, J. Horvath, J. R. Knickman, D. C. Colby, S. Schear, and M. Jung, “Chronic Conditions: Making the Case for Ongoing Care”, Accessed: Dec. 02, 2024. [Online]. Available: https://www.policyarchive.org/handle/10207/21756

3. J. Li and D. Porock, “Resident outcomes of person-centered care in long-term care: A narrative review of interventional research,” International Journal of Nursing Studies, vol. 51, no. 10, pp. 1395–1415, Oct. 2014. https://doi.org/10.1016/j.ijnurstu.2014.04.003.
4. M. Calvert, D. Kyte, G. Price, J. M. Valderas, and N. H. Hjollund, “Maximising the impact of patient reported outcome assessment for patients and society,” BMJ, vol. 364, p. k5267, Jan. 2019. https://doi.org/10.1136/bmj.k5267.
5. H. Nguyen, P. Butow, H. Dhillon, and P. Sundaresan, “A review of the barriers to using Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs) in routine cancer care,” Journal of Medical Radiation Sciences, vol. 68, no. 2, pp. 186–195, 2021. https://doi.org/10.1002/jmrs.421.
6. N. Black, “Patient reported outcome measures could help transform healthcare,” BMJ, vol. 346, p. f167, Jan. 2013. https://doi.org/10.1136/bmj.f167.
7. J. Greenhalgh, “The applications of PROs in clinical practice: what are they, do they work, and why?,” Qual Life Res, vol. 18, no. 1, pp. 115–123, Feb. 2009. https://doi.org/10.1007/s11136-008-9430-6.
8. C. F. Snyder, N. K. Aaronson, A. K. Choucair, T. E. Elliott, J. Greenhalgh, M. Y. Halyard, et al., “Implementing patient-reported outcomes assessment in clinical practice: a review of the options and considerations,” Qual Life Res, vol. 21, no. 8, pp. 1305–1314, Oct. 2012. https://doi.org/10.1007/s11136-011-0054-x.
9. V. Šafran, S. Lin, J. Nateqi, A. G. Martin, U. Smrke, U. Ariöz, et al., “Multilingual Framework for Risk Assessment and Symptom Tracking (MRAST),” Sensors, vol. 24, no. 4, Art. no. 4, Jan. 2024. https://doi.org/10.3390/s24041101.
10. I. Mlakar, U. Arioz, U. Smrke, N. Plohl, V. Šafran, and M. Rojc, “An End-to-End framework for extracting observable cues of depression from diary recordings,” Expert Systems with Applications, vol. 257, p. 125025, Dec. 2024. https://doi.org/10.1016/j.eswa.2024.125025.
11. B. Bardak and M. Tan, “Improving clinical outcome predictions using convolution over medical entities with multimodal learning,” Artificial Intelligence in Medicine, vol. 117, p. 102112, Jul. 2021. https://doi.org/10.1016/j.artmed.2021.102112.
12. N. Sasikumar and K. S. I. Mantri, “Transfer Learning for Low-Resource Clinical Named Entity Recognition,” in Proceedings of the 5th Clinical Natural Language Processing Workshop, T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, and A. Rumshisky, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 514–518. https://doi.org/10.18653/v1/2023.clinicalnlp-1.53.
13. A. Shaitarova, J. Zaghir, A. Lavelli, M. Krauthammer, and F. Rinaldi, “Exploring the Latest Highlights in Medical Natural Language Processing across Multiple Languages: A Survey,” Yearbook of Medical Informatics, vol. 32, pp. 230–243, Dec. 2023. https://doi.org/10.1055/s-0043-1768726.
14. S. Zhu, Supryadi, S. Xu, H. Sun, L. Pan, M. Cui, et al., “Multilingual Large Language Models: A Systematic Survey,” Nov. 19, 2024, arXiv: arXiv:2411.11072. https://doi.org/10.48550/arXiv.2411.11072.
15. J. Agerskov, K. Nielsen, and C. F. Pedersen, “Computationally Efficient Labeling of Cancer-Related Forum Posts by Non-clinical Text Information Retrieval,” SN COMPUT. SCI., vol. 4, no. 6, p. 711, Sep. 2023. https://doi.org/10.1007/s42979-023-02244-8.
16. O. Rohanian, M. Nouriborji, H. Jauncey, S. Kouchaki, I. C. C. Group, L. Clifton, et al., “Lightweight Transformers for Clinical Natural Language Processing,” Feb. 09, 2023, arXiv: arXiv:2302.04725. https://doi.org/10.48550/arXiv.2302.04725.
17. Y. Zhang, Y. Zhang, P. Qi, C. D. Manning, and C. P. Langlotz, “Biomedical and clinical English model packages for the Stanza Python NLP library,” Journal of the American Medical Informatics Association, vol. 28, no. 9, pp. 1892–1899, Sep. 2021. https://doi.org/10.1093/jamia/ocab090.
18. P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” Apr. 23, 2020, arXiv: arXiv:2003.07082. Accessed: May 08, 2024. [Online]. Available: http://arxiv.org/abs/2003.07082
19. J. Frei and F. Kramer, “GERNERMED -- An Open German Medical NER Model,” Dec. 10, 2021, arXiv: arXiv:2109.12104. https://doi.org/10.48550/arXiv.2109.12104.
20. C. Crema, T. M. Buonocore, S. Fostinelli, E. Parimbelli, F. Verde, C. Fundarò, et al., “Advancing Italian biomedical information extraction with transformers-based models: Methodological insights and

multicenter practical application," Journal of Biomedical Informatics, vol. 148, p. 104557, Dec. 2023. https://doi.org/10.1016/j.jbi.2023.104557.

21. L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, and A. Moreno-Sandoval, "A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine," BMC Medical Informatics and Decision Making, vol. 21, no. 1, p. 69, Feb. 2021. https://doi.org/10.1186/s12911-021-01395-z.
22. E. T. R. Schneider, J. V. A. de Souza, J. Knafou, L. E. S. e Oliveira, J. Copara, Y. B. Gumiel, et al., "BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition," in Proceedings of the 3rd Clinical Natural Language Processing Workshop, A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 65–72. https://doi.org/10.18653/v1/2020.clinicalnlp-1.7.
23. Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," J Am Med Inform Assoc, vol. 18, no. 5, pp. 552–556, 2011. https://doi.org/10.1136/amiajnl-2011-000203.
24. J. Tiedemann and S. Thottingal, "OPUS-MT – Building open translation services for the World," in Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada, Eds., Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 479–480. Accessed: Nov. 13, 2024. [Online]. Available: https://aclanthology.org/2020.eamt-1.61
25. Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, et al., "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning," Aug. 02, 2020, arXiv: arXiv:2008.00401. https://doi.org/10.48550/arXiv.2008.00401.
26. Z.-Y. Dou and G. Neubig, "Word Alignment by Fine-tuning Embeddings on Parallel Corpora," Aug. 12, 2021, arXiv: arXiv:2101.08231. https://doi.org/10.48550/arXiv.2101.08231.
27. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, arXiv: arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805.
28. L. A. Ramshaw and M. P. Marcus, "Text Chunking using Transformation-Based Learning," May 23, 1995, arXiv: arXiv:cmp-lg/9505040. https://doi.org/10.48550/arXiv.cmp-lg/9505040.
29. R. Nayak, "Focal Loss : A better alternative for Cross-Entropy," Medium. Accessed: Nov. 15, 2024. [Online]. Available: https://towardsdatascience.com/focal-loss-a-better-alternative-for-cross-entropy-1d073d92d075
30. A. Idrissi-Yaghir, A. Dada, H. Schäfer, K. Arzideh, G. Baldini, J. Trienes, et al., "Comprehensive Study on German Language Models for Clinical and Biomedical Text Understanding," in Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds., Torino, Italia: ELRA and ICCL, May 2024, pp. 3654–3665. Accessed: Dec. 10, 2024. [Online]. Available: https://aclanthology.org/2024.lrec-main.324

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.