International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Enhancing Vietnamese Medical Named Entity Recognition with BERT and CRF

Huy Huynh[1], Thanh Cao[1], Thanh Minh Cao[1], Hai Tran[2*]

Submitted:14/03/2024 **Revised**: 29/04/2024 **Accepted**: 06/05/2024

Abstract: Named Entity Recognition (NER) is used in the medical field, and this study looks into how it might be used to extract important information from unstructured textual data. For a variety of medical applications, it is critical to accurately identify things such as illnesses, available treatments, and healthcare initiatives. This paper presents a new method to improve NER performance in the medical domain by using Conditional Random Fields (CRF) in conjunction with Bidirectional Encoder Representations from Transformers (BERT). Superior performance and accuracy in identifying medical entities are achieved by the BERT-CRF model, which combines the contextual awareness of BERT with the sequence modeling capabilities of CRF. We test different BERT iterations in this work, including Base-BERT, RoBERTa, and XLM-RoBERTa. We used two distinct datasets for our tests. Two datasets are available: one in English with writings that have complicated and ambiguous semantics, and the other in Vietnamese with medical texts gathered from the Ministry of Health's Electronic Portal in Vietnam. The outcomes show that NER performed admirably on both datasets, especially in the medical field. The model using the XLM-RoBERTa version gives the best results for the Vietnamese medical dataset. This indicates how well NER for medical entity extraction may be improved in terms of accuracy and stability by integrating XLM-RoBERTa-CRF. The results of this study improve the suitability of natural language processing techniques for use in the medical field and have broad potential applications.

Keywords: *named entitye recognition , vietnamese healthcare entity, deep learning, xlm-roberta, crf*

1. Introduction

In the field of natural language processing (NLP), Named Entity Recognition (NER) stands as one of the primary tasks in text analysis and understanding. The task of NER is to identify and classify named entities within text into different categories, including entities such as persons, locations, organizations, time expressions, and various others. Executing NER not only aids computers in understanding the structure of text but also opens up numerous opportunities for information extraction. Through the recognition and classification of named entities, computers can comprehend the relationships between entities, thereby constructing a structured database to support information retrieval, data analysis, and decision-making processes. The applications of NER can be found in various fields, including knowledge management, finance, and healthcare, in addition to the topic of natural language processing[1][2][3]. Several well-known NER applications include: Information Extraction [4][5][6]. Question Answering Systems [7][8], Sentiment Analysis [9][10], Document Classification [11][12].

Significant developments in Artificial Intelligence (AI) have spurred innovation in a number of sectors, including medical, in recent years [13]. AI has been useful in improving disease management, forecasting results, and improving diagnostics in the medical field. Particularly, there is a great deal of promise for medical improvement with the use of novel AI approaches in picture categorization from imaging modalities like X-rays [14] and dermatological [15]. Meanwhile, as AI techniques have advanced, there has been a rise in interest in Natural Language Processing, or NLP, in the medical field. Enhancing search capabilities, assessing evidence-based procedures, and creating intelligent healthcare applications are all made possible by using natural language processing (NLP) to analyze and synthesize information from medical texts. The developments in AI and NLP not only open up new avenues for medical practice, but they also greatly improve patient outcomes and the quality of healthcare delivery [16].

The fact that words in the medical profession are not conventional nouns and sometimes have complex grammatical structures is one of the main obstacles. Units of measurement, clinical expressions, specialist medical terminology, and other complex characteristics are examples of entities. In order to maintain the accuracy and dependability of the identification process, NER models must contend with an environment that is constantly updated with new knowledge about illnesses,

[1]*Faculty of Information Technology, Saigon University (SGU), HCM, Vietnam*
[2]*Department of Information Technology, Ho Chi Minh University of Education (HCMUE), HCM, Vietnam*
hnkhuy@sgu.edu.vn
thanh.cao@sgu.edu.vn
**Corresponding author: haits@hcmue.edu.vn*

medical advancements, and therapeutic approaches [17,18]. We illustrate some examples of complex entities in the medical field in Table I.

Table I. Some examples of complex entities in the medical field

Type of Entities	Content	Comment
Disease name	Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS) Acute Respiratory Tract Infection caused by viruses	These disease names often have complex and diverse structures
Names of medical organizations	-Centers for Disease Control and Prevention of the United States(CDC): *This is an important public health organization in the United States, responsible for research, surveillance, and disease prevention.* -National Institute of Hematology - Blood Transfusion: *Specializing in research and providing services in hematology and blood transfusion for patients.*	These organizations often have complex and contextually and semantically diverse names, making identification and classification by named entity recognition models difficult.
Names of medical programs	-Free Prevention and Treatment Program for the Poor: *A government program aimed at providing free basic healthcare services to the poor and minority ethnic groups.* -Nutrition and Health Enhancement Program for Children: *Providing nutrition and healthcare services for children to aid in their development and disease prevention*	These programs often have broad goals and scope, and their names can be complex and not easily identifiable by named entity recognition models.
Treatments	-Electroencephalogram (EEG) examination: *An imaging diagnostic method is used to record brain electrical activity to identify brain disorders.* -Laparoscopic Surgery: *A minimally invasive surgical method, often utilized for appendectomy or addressing gastrointestinal issues.*	These treatment methods often involve specialized terminology and complex language structures.

As such, improving NER models' capacity to handle various and complicated named entities in the medical domain is a significant and exciting area of research for the NLP and AI communities. We propose to use Pre-trained Language Models (PLMs) rigorously to detect such entities. By using PLMs, you can reduce the amount of time needed for training and testing procedures while also improving system performance. In order to determine the best technique for entity recognition in medical texts, we experimented with different PLMs in this study and assessed the outcomes.

2. Background and Related Work

Named Entity Recognition (NER) has grown in importance and interest as a study issue in recent natural language processing (NLP) studies. The identification and classification of named entities in text is the mission of NER, as seen by the variety of approaches and technologies used to address this issue. Prominent research works have tackled the application of conventional machine learning methods such Conditional Random Field (CRF) [19], Support Vector Machine (SVM) [20], and Perceptrons [21]. These techniques have made a significant contribution to NER's solution. These studies also show, however, that they frequently lack the adaptability necessary to deal with opaque and complicated structured entities, particularly in specialized industries like banking, law, or medicine.

BiLSTM-CRF [22][23], BERT [24][25], and other deep learning models can now be used thanks to the development of machine learning and deep learning. Owing to their capacity to pick up on intricate details and contextual relationships, these models have demonstrated outstanding performance in NER and natural language processing. In particular, the BERT model (Bidirectional Encoder Representations from Transformers) has attracted the attention of the academic community and shown impressive performance on common datasets like CoNLL03/OntoNotes. With BERT's initial release, numerous modifications and enhancements have been made to improve its effectiveness and adaptability to a variety of activities and languages.

- Base-BERT: Base-BERT is the basic version of the BERT model, trained on a large amount of English language data. It consists of 12 Transformer layers and 768 hidden representation dimensions [26].

- RoBERTa (Robustly optimized BERT approach): RoBERTa is a variant of BERT developed by Facebook AI through fine-tuning the training process and changing how the data is processed. RoBERTa improves the performance of BERT by using additional data and new optimization techniques [27].

- XLM-RoBERTa (Cross-lingual Language Model RoBERTa): XLM-RoBERTa is a variant of RoBERTa developed by Facebook AI for multilingual language processing. It is trained on a large amount of data from various languages, enhancing the model's transferability across different languages, including Vietnamese language [28].

These BERT variations have proven to be quite effective on a variety of NLP tasks and have grown to be indispensable resources for both academic study and practical use. But there are still a lot of obstacles in the way of NER development. More specifically, managing novel, diverse, and non-fixed sorts of entities and automatically recognizing and classifying these entities within a complex and diverse linguistic environment are major issues. This problem stems from a number of factors, one of which is that NER models frequently make predictions based only on surface-level entity patterns, neglecting contextual evidence. Consequently, when out-of-distribution (OOD) entity patterns are added, even the most advanced NER models perform poorly in generalizing to out-of-domain circumstances [29]. Furthermore, the primary problem is the tainted label quality as a result of intrinsic flaws including positive type errors, false negatives, and false positives [30]. Additionally, entity detection of subjects within specialist professions becomes considerably more difficult and challenging if the research focuses on specific domains like finance, law, or medical.

Numerous studies have made a substantial contribution to the advancement of techniques and technology for medical entity recognition, particularly in the healthcare sector, creating new avenues for the use of deep learning and machine learning in this field. Machine learning algorithms and neural networks are employed in certain medical entity recognition technologies. In order to satisfy the need for medical entity recognition in the healthcare industry, these models frequently rely on conventional methods or their improvements.

- MedTag: MedTag is a NER model built on the basis of a Multi-task Recurrent Neural Network. This model is trained on medical data to recognize entities such as disease names, medication names, and medical information in text [31].

- N2C2: The N2C2 model is a Named Entity Recognition (NER) model developed for the Named Entity Recognition for Clinical Records (N2C2) competition. This model utilizes a complex neural network architecture to identify medical entities within clinical notes and electronic health records [32].

- CRF-BiLSTM: This model combines a Multi-task Recurrent Neural Network (RNN) and a Conditional Random Field (CRF) to recognize medical entities within medical text [33].

- SVM-Based Models: Support Vector Machine (SVM)-based models are also used in medical entity recognition, where features of words and context are utilized to train a classifier to identify medical entities [34].

Additionally, there are notable models applying BERT for medical entity recognition, which are significant research contributions in this field.

- BioBERT: is a variant of BERT trained on large medical data. This research contributes to improving the performance of medical entity recognition on medical datasets and represents a significant advancement in applying deep learning models for natural language processing in the medical domain [35].

- ClinicalBERT: is a variant of BERT trained on clinical notes data. This research emphasizes the use of deep learning models for analyzing and predicting patient readmissions based on clinical notes [36].

- SciBERT: is a variant of BERT fine-tuned for applications in the scientific and medical domains. This model is trained on data from scientific papers to enhance its understanding and language processing capabilities in the scientific and medical fields [37].

- BERT-BiLSTM-CRF: combines the strengths of BERT in contextual understanding, the sequence processing capability of BiLSTM, and the ability of CRF to model relationships between labels, thus improving the performance of the model in entity recognition tasks [38].

In fact, because of the complexity and diversity of medical terminology as well as the specialization and originality of medical terms, medical entity recognition in practice faces distinct obstacles.

- Difficult domain-specific language: The medical industry uses a lot of specialist terms, expressions, and acronyms that are difficult for non-medical professionals to comprehend. One major problem for entity recognition algorithms is the diversity of medical language.

- Different types of entities: Test results, hospital names, medication names, disease names, and more are all included in the field of healthcare. A system for recognizing entities must be able to handle a variety of entity kinds due to this diversity.

- Scalability and ongoing updating: New terminology, approaches to therapy, and medical advancements are all part of the ever-evolving field of medicine. In order to satisfy evolving demands, a medical entity recognition system needs to be able to grow and update on a regular basis.

- Data heterogeneity: Electronic health records, medical literature, online forums, and other sources of medical data can all display a variety of formats and structural variations. The process of creating models for medical entity recognition is made more difficult by this heterogeneity.

Acquiring a thorough grasp of medical terminology in addition to sophisticated natural language processing methods is necessary to identify medical entities and overcome the unique obstacles this domain presents. Consequently, in order to improve productivity and shorten training and testing times, deep learning techniques, big data, and natural language processing must be combined. Pre-trained Language Models (PLMs) must also be fully employed. Potential for improving NER and its applications in real-world domains like legal, finance, and medical can be realized through evaluating outcomes across several PLMs. Our study methodology also follows this path.

3. Vietnamese Healthcare Entity Recognition proposed model (viHER)

In this study, we construct an architecture named BERT-CRF by taking the standard model BERT-BiLSTM-CRF and deleting its BiLSTM layer [38]. Utilizing the benefits of deep learning-based language models, this approach extracts critical information for the medical domain. Our BERT-CRF model simplifies the architecture and makes it easier to deploy and train by feeding the word representation vectors produced by BERT directly into the Conditional Random Field (CRF) layer for efficient named entity recognition. Additionally, using CRF to model the relationship between labels improves entity recognition performance [33]. Subsequently, the model will be assessed by testing on medical dataset.

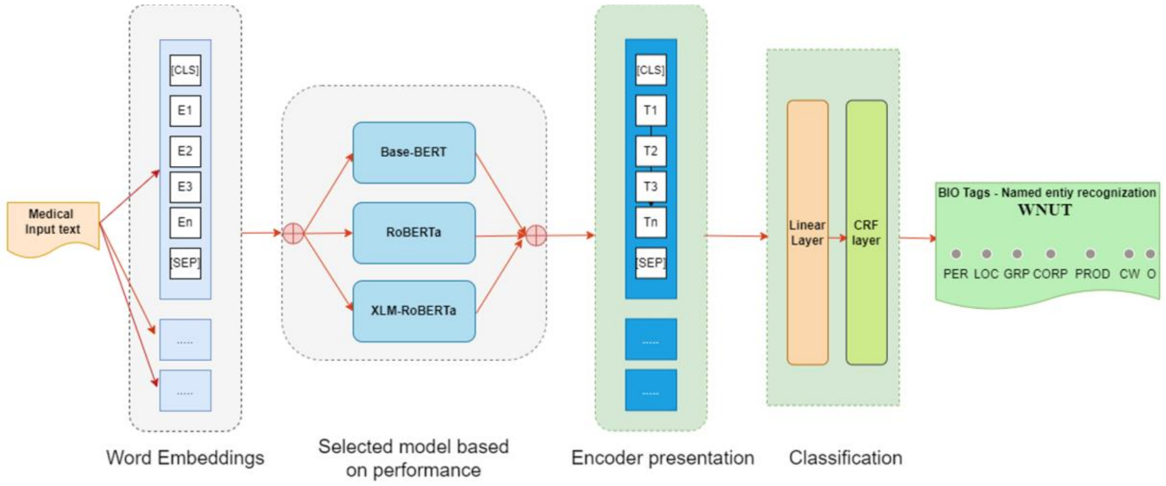

Fig 1. Proposed model for named entity recognition (viHER)

BERT is a pre-trained language model in this model that was trained on a substantial quantity of unlabeled text data. To learn word and phrase representations in natural languages, it makes use of a transformer architecture. We will test on multiple versions of BERT in this model to evaluate the entity recognition performance, including **Base-BERT, RoBERTa, and XLM-RoBERTa**. One of the great benefits of BERT is its ability to understand the context from both sides of the word, helping to improve the semantic representation of words and sentences [26].

CRF (Conditional Random Field): CRF is a type of sequential probability model commonly used in entity recognition tasks. It is capable of modeling dependencies between consecutive entities in a data sequence. CRF is often used after a neural network to combine information from the neural network with information about the chain structure [19].

The way the model works is as follows: for a sentence the input text will be represented by vectors from the BERT model, then calculate scores for each entity label for each word in the sentence. These scores are typically calculated using a neural network (a fully connected layer) applied to the word's representation from the BERT model. Next, we use the CRF model to calculate the probability of an entity label sequence for the entire sentence, based on the scores calculated from BERT and the sequential learning model constraints of CRF. This helps improve tag string prediction by considering the global context of the sentence instead of just considering the context of individual words. To calculate the probability of a label sequence from CRF, we use a softmax function and a loss function with parameters updated through the backpropagation algorithm during model training. Combining the advantages of sequential learning modeling capabilities of CRF with context representation from the BERT model results in an efficient way to tackle the entity recognition challenge. This improves prediction accuracy and consistency [33][36].

In this study, we utilized the HuggingFace Transformers library, which is considered one of the most powerful and popular PyTorch interfaces for handling BERT models. Subsequently, we conducted experiments with several BERT-based models downloaded from HuggingFace, including BERT, RoBERTa, and XLM-RoBERTa with different versions. Due to the necessity of training a large neural network, we chose Google Colab to leverage the GPU and TPU resources available in this environment. Regarding hyperparameter tuning, we experimented with various values around those recommended as "working well across all tasks" [26], and here are the parameter sets that yielded the best results.: *Batch size: 64, Learning rate: 2e-5, Number of epochs: 10, Maximum sequence length: 128.*

We experimented on two datasets in order to assess the model:

- An *English dataset* spanning various domains.

- A *Vietnamese dataset* collected from the portal of the Vietnamese Ministry of Health.

For the English dataset, we evaluated the BERT-CRF model on three versions of BERT: BERT, RoBERTa, and XLM-RoBERTa. Next, based on the results from this stage and understanding the characteristics of each BERT version, where XLM-RoBERTa is a pretrained model trained on multiple languages including Vietnamese *(24757M tokens, size 137.3 GiB data Vietnamese on traing dataset)* [28]. Therefore, we selected the **XLM-RoBERTa-CRF** model to experiment on the Vietnamese healthcare dataset. We have called this model **viHER** (**Vi**etnamese **H**ealthcare **E**ntity **R**ecognition).

4. Experimental and Discussion

Additionally, there are 2 characters, B and I, where B represents the beginning of an entity name, and I represents all the words inside an entity and all the words outside an entity.

Therefore, there are a total of 13 tags: B-PER, I-PER, B-LOC, I-LOC, B-GRP, I-GRP, B-CORP, I-CORP, B-PROD, I-PROD, B-CW, I-CW, and O. The data will be formatted accordingly CoNLL[1]:

Table II. Entity types used in this study

Entity	**Original name**	**Meaning**
PER	person	Name of person
LOC	location	Name of place, location
CORP	corporation	Company, corporation
GRP	group	Organizations and groups
PROD	product	Product
CW	creative-work	Names of programs and activities

In this study, we will focus on 6 types of entities according to WNUT 2017 :

Table III. Illustration of English and Vietnamese dataset formats

English	**Entity**	**Vietnamese**	**Entity**
kingdom	B_CW	Buổi	B-CW
hospital	I-CW	hiến	I-CW
,	O	máu	I-CW
lewiston	B-LOC	của	O
from	O	Hội	B-GRP
stephen	B-PER	Chữ	I-GRP
king	I-PER	thập	I-GRP
of	O	đỏ	I-GRP
the	O	Mỹ	I-GRP
same	O	tại	O
name	O	26	O
		bang	O
		và	O
		Washington	B-LOC
		DC	I-LOC

4.1. MultiCoNER dataset

We will use the MultiCoNER dataset provided in to train the model and test its performance with the English language. Table IV,V lists the sizes of the data sets in the study.

[1] https://universaldependencies.org/docs/format.html

Table IV. Total amount of training and testing data

Training	Practice phase
15,300	800

Table V. Details the amount of data for each entity type

PER		LOC		GRP		CORP		CW		PROD	
Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
5,397	290	4,799	234	3,571	190	3,111	193	3,752	176	2,923	147

4.2.

Dataset collected from the information portal of the Ministry of Health of Vietnam

To evaluate the effectiveness of the model, we constructed a dataset of Vietnamese texts in the medical field. The data was collected from the Ministry of Health of Vietnam's portal[2] . The official portal website of the Ministry of Health of Vietnam is a primary source of information managed and provided by the Government of Vietnam's Ministry of Health. This website mainly offers information about activities, policies, and services related to the healthcare sector domestically. Some basic information that can be found on the portal website of the Ministry of Health of Vietnam includes:

- Information on healthcare policies: The website provides information on healthcare policies, regulations, and laws related to the healthcare sector in Vietnam. This is where the community can look up and understand the legal regulations and implementation guidelines in the field of healthcare.

- Healthcare news and events: The website offers the latest news on events, activities, and healthcare programs of the Ministry of Health. This information helps the community stay updated on the healthcare situation, community healthcare campaigns, and other important healthcare events.

- Public healthcare services: The website provides information on public healthcare services provided by the Ministry of Health and other healthcare units nationwide. This information includes vaccination programs, disease control, and other community healthcare services.

Health information lookup: The website offers tools for looking up healthcare information such as lists of healthcare facilities, information on various diseases, medication usage guidelines, and other healthcare advice.

- Contact and feedback: The website provides information on how to contact the Ministry of Health, addresses, and contact information of agencies and units under the Ministry of Health. Additionally, there is information on how to provide feedback and suggestions on healthcare issues.

The data is divided into 100 test cases, each corresponding to 100 unlabeled files containing some medical information collected for a specific purpose (containing types of entities to be identified). The data format is also structured according to the CoNLL3 standard, table VI

We will use this dataset to test the model's entity recognition effectiveness with Vietnamese documents in the medical field.

4.3. Measure

To evaluate the model, we will use the Precision, Recall, and F1-score metrics. Specifically:

- Precision is defined as the ratio of the number of true positive predictions made by the model to the total number of predictions made by the model that are positive.

- Recall is defined as the ratio of the number of true positive predictions made by the model to the total number of actual positives (or the total number of points labeled as positive initially).

However, Precision or Recall alone does not provide a comprehensive assessment of model quality.

- F1-score is calculated according to the formula:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

[2] https://moh.gov.vn/

[3] https://universaldependencies.org/docs/format.html

$$F1 - Score = \frac{2 \, * \, \text{Precision} \, * \, \text{Recall}}{\text{Precision} \, + \, \text{Recall}}$$

4.4. Results on dataset MultiCoNER

For the testing phase, evaluating the effectiveness with different BERT models, we use the MultiCoNER data set, then in turn test the BERT-CRF model with different versions of BERT: base-bert, bert-large, roberta, xlm-roberta. The following results:

Based on this result, it can be seen that the model using two pre-trained models, bert-large and xlm-roberta, gives slightly better results than the other two models. With xlm-roberta giving the best results for the three entity types PER, CW, GRP, while bert-large gives the best results for the two entity types LOC, CORP and base-bert gives the best results for the PROD type.

Continuing with the results of Table 5, it can be seen that the model using xlm-roberta has better results than the remaining models when calculated according to the measures, followed by bert-large.

However, in general, the results between models are not much different, so to improve efficiency in NER, in addition to using pre-trained models, it is necessary to learn other methods to achieve results. best results. When analyzing each measure, the BERT model with the bert-large version gives the highest Precison result (0.849), showing that the BERT model has high accuracy in labeling entities. While xlm-roberta gives the best results in two values: Recall (0.882) and F1-Score (0.863).

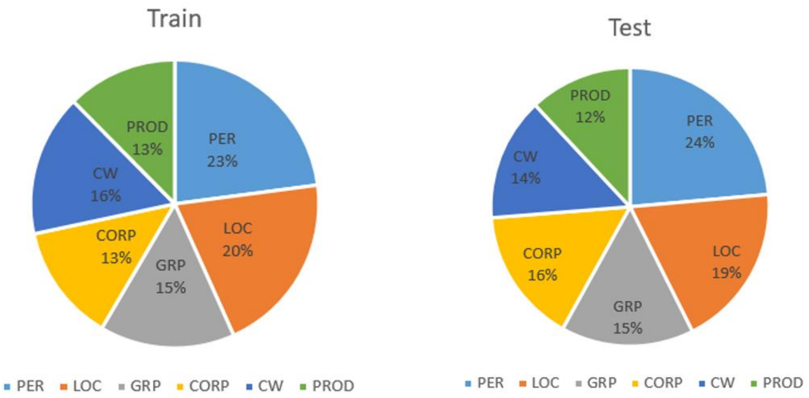

Fig 2. Percentage of training and test data

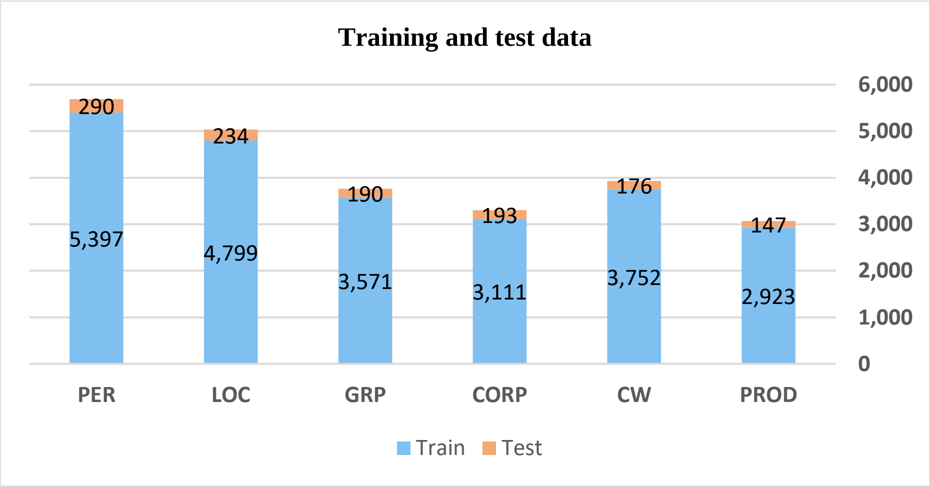

Fig 3. Illustration of training and test data

Table VI. Some illustrations of medical field test data

PER	**LOC**	**GRP**	**CORP**	**PROD**	**CW**
Đại_ _ B-PER	huyện_ _ B-LOC	Trung_ _ B-GRP	Bảo_ _ B-CORP	Thuốc_ _ B-PROD	chiến_ _ B-CW
danh_ _ I-PER	Ea_ _ I-LOC	tâm_ _ I-GRP	hiểm_ _ I-CORP	hóa_ _ I-PROD	dịch_ _ I-CW
y_ _ I-PER	Kar_ _ I-LOC	Kiểm_ _ I-GRP	Xã_ _ I-CORP	dược_ _ I-PROD	nhắn_ _ I-CW
Hải_ _ I-PER	và_ _ O	soát_ _ I-GRP	hội_ _ I-CORP	sinh_ _ O	tin_ _ I-CW

Thượng_ _ I-PER	TP_ _ O	bệnh_ _ I-GRP	Việt_ _ I-CORP	phẩm_ _ O	ủng_ _ I-CW
Lãn_ _ I-PER	._ _ O	tật_ _ I-GRP	Nam_ _ I-CORP	y_ _ O	hộ_ _ I-CW
Ông_ _ I-PER	Buôn_ _ B-LOC	tỉnh_ _ I-GRP		tế_ _ O	bệnh_ _ I-CW
Lê_ _ I-PER	Ma_ _ I-LOC				nhân_ _ I-CW
Hữu_ _ I-PER	Thuột_ _ I-LOC				nghèo_ _ I-CW
Trác_ _ I-PER					

The data is distributed as follows:

Table VII. Number of each entity type

Type of entities	**PER**	**LOC**	**GRP**	**CORP**	**PROD**	**CW**
Number	88	126	388	20	40	161

4.5. Results on dataset collected from Vietnam Ministry of Health information portal

Based on the test results in the above step, we chose the **XLM-RoBERTa** version for testing on real application data from the Vietnam Ministry of Health portal to evaluate the level of recognition of named entities of model with Vietnamese data in the medical field.

The results show that two groups of entities PER (proper name of person) and LOC (name of place, location) are recognized quite accurately (0.93 and 0.94) on both Vietnamese data, including similar data relatively complicated.

The system recognized the entity types GRP, CORP and CW at a fairly accurate level of 0.84, 084 and 0.8 respectively, proving that the system has recognized entities that are quite complex and have unclear semantics (depending on the specific context).

The entity type PROD gives the lowest result of 0.66, this is the name of diseases and medicinal herbs. Because most of these words will have many semantics in different contexts, considering it in the medical field will give error, for example an error like below:

Thuốc_ _ O → *wrong result*

hóa_ _ O → *wrong result*

dược_ _ O → *wrong result*

sinh_ _ B-CORP

phẩm_ _ I-CORP

y_ _ I-CORP

tế_ _ I-CORP

Table VIII. Evaluation results of tested models with 6 types of entities

Model	**PER**			**LOC**			**CW**			**GRP**			**CORP**			**PROD**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
base-bert	0.92	0.96	0.94	0.82	0.89	0.86	0.67	0.76	0.71	0.83	0.88	0.85	0.86	0.80	0.83	0.66	**0.86**	**0.75**
bert-large	**0.96**	0.97	0.96	0.85	**0.91**	**0.88**	0.74	0.79	0.77	0.85	**0.94**	0.89	**0.90**	**0.83**	**0.87**	**0.71**	0.74	0.72
roberta	0.91	0.90	0.90	0.86	0.89	**0.88**	0.67	0.72	0.69	0.79	0.85	0.82	0.88	0.80	0.84	0.69	0.78	0.73
xlm-roberta	**0.96**	**0.98**	**0.97**	**0.86**	0.89	0.87	**0.77**	**0.83**	**0.80**	**0.87**	0.91	**0.89**	0.86	0.82	0.84	0.67	0.79	0.73

Table IX. Results on dataset MultiCoNER

Model	Precision	Recall	F1-score
base-bert	0.804	0.869	0.835
bert-large	0.849	0.876	0.863
roberta	0.812	0.832	0.822
xlm-roberta	**0.845**	**0.882**	**0.863**

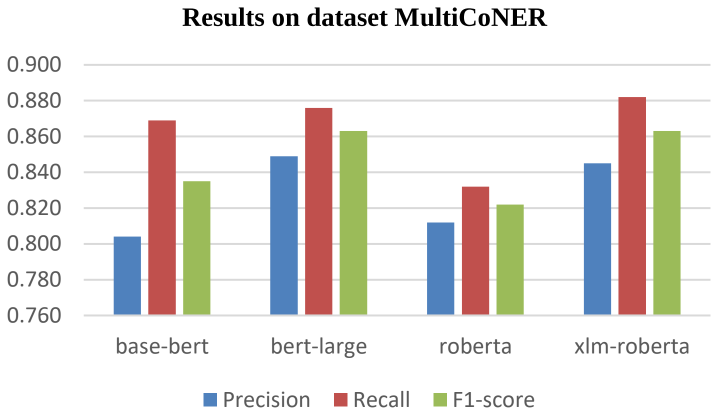

Fig 4. Results on dataset MultiCoNER

Table X. Performance of the model on medical data set according to 6 entity types

Type of entities	PER	LOC	GRP	CORP	PROD	CW
Accuracy rate	0.93	0.94	0.84	0.84	0.66	0.80

5. Conclusion

Our work introduces a Named Entity Recognition (NER) system designed to process datasets with complex and ambiguous semantic features, such as those from the medical field. The objective of the project is to create a dependable NER system that can swiftly and accurately identify and classify named things in medical texts, regardless of how complex the context or how ambiguous the data is. Our method combines Named Entity Recognition (NER) task performance with the use of a BERT-CRF model. First, the input data is preprocessed and fed into the BERT model with various variants to evaluate performance, including Base-BERT, RoBERTa, and XLM-RoBERTa. Next, the output results from the BERT model are passed through a linear layer and finally a CRF layer is applied to identify the entities. Additionally, we also conduct experiments on some hyperparameter adjustments to optimize the model's performance.

In order to assess the model's performance and guarantee the system's flexibility and dependability, we ran tests on two primary datasets: the MultiCoNER dataset and a self-constructed healthcare dataset gathered from the Vietnam Ministry of Health Portal. Our suggested NER approach produced encouraging results on both datasets, as shown by the experimental results. Especially, the XLM-RoBERTa-CRF model yields high performance in healthcare entity recognition. This suggests that the model not only possesses strong generalization abilities across various datasets but also offers a strong basis for future processing and analysis of healthcare data.

In the future, we expect to continue experimenting with various Pre-trained Language Models (PLMs) to optimize the performance of the system. Additionally, we plan to expand the scope of research by applying deep learning techniques and advanced language theories to enhance the accuracy of entity recognition, especially in processing complex entities, particularly in the medical field.

References

[1] Bartolini, I., Moscato, V., Postiglione, M., Sperlì, G., & Vignali, A. (2023). Data augmentation via context similarity: An application to biomedical Named Entity Recognition. Information Systems, 119, 102291.

[2] Sabane, M., Ranade, A., Litake, O., Patil, P., Joshi, R., & Kadam, D. (2023, May). Enhancing Low Resource NER using Assisting Language and Transfer Learning. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 1666-1671). IEEE.

[3] Jarrar, M., Abdul-Mageed, M., Khalilia, M., Talafha, B., Elmadany, A., Hamad, N., & Omar, A. (2023). WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. arXiv preprint arXiv:2310.16153.

[4] Ariyanto, A. D. P., Fatichah, C., & Purwitasari, D. (2023, July). Semantic Role Labeling for Information Extraction on Indonesian Texts: A Literature Review. In 2023 International Seminar on Intelligent Technology and Its Applications (ISITIA) (pp. 119-124). IEEE.

[5] Rattanatamrong, P., Boonpalit, Y., & Boonnavasin, M. (2024). Utilising crowdsourcing and text mining to enhance information extraction from social media: A case study in handling COVID-19 supply requests in Thailand. Journal of Information Science, 01655515231220164.

[6] Al Mamun, M. A., Azad, M. A. K., & Pramanik, M. I. (2023). Crime-Finder: A System for Extraction and Visualization of Crime Data from Bengali Online Newspaper Articles. In Applied Informatics for Industry 4.0 (pp. 97-108). Chapman and Hall/CRC.

[7] Younes, Y., & Scherp, A. (2023). Question Answering versus Named Entity Recognition for Extracting Unknown Datasets. IEEE Access.

[8] Kruff, A., & Tran, A. H. (2023, July). Billie-newman at semeval-2023 task 5: Clickbait classification and question answering with pre-trained language models, named entity recognition and rule-based approaches. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 1542-1550).

[9] Vági, R. (2023). How Could Semantic Processing and Other NLP Tools Improve Online Legal Databases?. TalTech Journal of European Studies, 13(2), 138-151.

[10] Shafqat, S., Anwar, Z., Javaid, Q., & Ahmad, H. F. (2023). NER Sequence Embedding of Unified Medical Corpora to Incorporate Semantic Intelligence in Big Data Healthcare Diagnostics. Qeios.

[11] Szczepanek, R. (2023). A deep learning model of spatial distance and named entity recognition (SD-NER) for flood mark text classification. Water, 15(6), 1197.

[12] Kutbi, M. (2023). Named Entity Recognition Utilized to Enhance Text Classification while Preserving Privacy. IEEE Access.

[13] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.

[14] Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. (2018). Artificial intelligence in radiology. Nature Reviews Cancer, 18(8), 500-510.

[15] Du-Harpur, X., Watt, F. M., Luscombe, N. M., & Lynch, M. D. (2020). What is AI? Applications of artificial intelligence to dermatology. British Journal of Dermatology, 183(3), 423-430.

[16] Navarro, D. F., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., & Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. International Journal of Medical Informatics, 105122.

[17] Li, M., & Zhang, R. (2023). How far is language model from 100% few-shot named entity recognition in medical domain. arXiv preprint arXiv:2307.00186.

[18] Miah, M. S. U., Sulaiman, J., Sarwar, T. B., Islam, S. S., Rahman, M., & Haque, M. S. (2023, July). Medical named entity recognition (medner): A deep learning model for recognizing medical entities (drug, disease) from scientific texts. In IEEE EUROCON 2023-20th International Conference on Smart Technologies (pp. 158-162). IEEE.

[19] Das, A., & Garain, U. (2014). Crf-based named entity recognition@ icon 2013. arXiv preprint arXiv:1409.8008.

[20] Ekbal, A., & Bandyopadhyay, S. (2010). Named entity recognition using support vector machine: A language independent approach. International Journal of Electrical and Computer Engineering, 4(3), 589-604.

[21] Higashiyama, S., Mathieu, B., Seki, K., & Uehara, K. (2015). Cost-sensitive structured perceptron incorporating category hierarchy for named entity

recognition. Journal of Information and Communication Technology, 14, 1-20.

[22] Jimmy, L., Nongmeikappam, K., & Naskar, S. K. (2023). Bilstm-crf Manipuri ner with character-level word representation. Arabian Journal for Science and Engineering, 48(2), 1715-1734.

[23] Arslan, S. (2024). Application of BiLSTM-CRF model with different embeddings for product name extraction in unstructured Turkish text. Neural Computing and Applications, 1-12.

[24] Guo, B., & Liu, H. (2023). Integration of natural and deep artificial cognitive models in medical images: BERT-based NER and relation extraction for electronic medical records. Frontiers in Neuroscience, 17, 1266771.

[25] Song, Z., Xu, W., Liu, Z., Chen, L., & Su, H. (2023, August). A BERT-Based Named Entity Recognition Method of Warm Disease in Traditional Chinese Medicine. In 2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA) (pp. 1226-1231). IEEE.

[26] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[27] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[28] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[29] Ma, R., Wang, X., Zhou, X., Zhang, Q., & Huang, X. J. (2023, December). Towards Building More Robust NER datasets: An Empirical Study on NER Dataset Bias from a Dataset Difficulty View. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 4616-4630).

[30] Li, Y., Zhou, K., Qiao, Q., Wang, Q., & Li, Q. (2024). Re-Examine Distantly Supervised NER: A New Benchmark and a Simple Approach. arXiv preprint arXiv:2402.14948.

[31] Giachelle, F., Irrera, O., & Silvello, G. (2021). MedTAG: a portable and customizable annotation tool for biomedical documents. BMC Medical Informatics and Decision Making, 21(1), 352.

[32] Mahajan, D., Liang, J. J., Tsou, C. H., & Uzuner, Ö. (2023). Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. Journal of Biomedical Informatics, 144, 104432.

[33] Jingru, L., Yang, S., Rui, J., Yipeng, Z., Yong, L., & Jingdong, M. (2020). A BiLSTM-CRF Model for Protected Health Information in Chinese. Data Analysis and Knowledge Discovery, 4(10), 124-133.

[34] Hamad, R. M., & Abushaala, A. M. (2023, May). Medical Named Entity Recognition in Arabic Text using SVM. In 2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA) (pp. 200-205). IEEE.

[35] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.

[36] Huang, K., Altosaar, J., & Ranganath, R. (2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342.

[37] Cevik, M., Mohammad Jafari, S., Myers, M., & Yildirim, S. (2023). Sequence Labeling for Disambiguating Medical Abbreviations. Journal of Healthcare Informatics Research, 7(4), 501-526.

[38] Hanqing, Z. H. A. O., Li, Y., & Zhang, S. (2024). Symptom Extraction of Internal Medicine Diseases of Traditional Chinese Medicine Based on BERT-BiLSTM-CRF Model.