*Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 328–337
Marseille, 20-25 June 2022


# ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining

**Minh Phuc Nguyen** [†,*], **Vu Hoang Tran** [†,*], **Vu Hoang** [†], **Ta Duc Huy** [†]
**Trung H. Bui**, **Steven Q. H. Truong** [†]
[†]VinBrain
{v.minhng,v.vutran,v.vuhoang,v,huyta,v.brain01}@vinbrain.net
bhtrung@gmail.com

**Abstract**

Pre-trained language models have become crucial to achieving competitive results across many Natural Language Processing (NLP) problems. For monolingual pre-trained models in low-resource languages, the quantity has been significantly increased. However, most of them relate to the general domain, and there are limited strong baseline language models for domain-specific. We introduce **ViHealthBERT**, the first domain-specific pre-trained language model for Vietnamese healthcare. The performance of our model shows strong results while outperforming the general domain language models in all health-related datasets. Moreover, we also present Vietnamese datasets for the healthcare domain for two tasks: Acronym Disambiguation (AD) and Frequently Asked Questions (FAQ) Summarization. We release ViHealthBERT to facilitate future research and downstream applications for Vietnamese NLP in domain-specific. Our dataset and code are available in https://github.com/demdecuong/vihealthbert.

**Keywords:** Low-resource language, language model, healthcare, acronym disambiguation, summarization
* denotes equal contribution

## 1. Introduction

Recent large-scale language models show remarkable achievements in key NLP tasks such as Question Answering (Devlin et al., 2019) and Text Summarization (Raffel et al., 2020). More research studies have discovered multilingual language models for better performance in a wide range of downstream tasks without considering its language, and lots of low-resource languages benefit from this. (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Lan et al., 2020). However, in low-resource languages, monolingual language models still demonstrate their superiority on general domain benchmarks (Wu and Dredze, 2020). Leveraging mBERT(Devlin et al., 2019) and ELECTRA(Clark et al., 2020) method, viBERT and vELECTRA(Bui et al., 2020) have been proposed and showed superior performance in Vietnamese Named Entity Recognition (NER) and POS Tagging tasks. Then PhoBERT (Nguyen and Tuan Nguyen, 2020) introduces the first large-scale BERT-based Vietnamese language model and becomes a strong baseline for Vietnamese NLP tasks, which utilizes a high-volume corpus of news following RoBERTa(Liu et al., 2019) training strategy.

While domain-specific language models are abundant (Lee et al., 2020; Rasmy et al., 2021; Roy and Pan, 2021; Gu et al., 2021), studies addressing Vietnamese domain-specific tasks are few and far between, as existing Vietnamese language model studies still focus on the general domain. The first and the only is a language model for legal-domain (Chau et al., 2020) that applies for Vietnamese legal text retrieval, but the study meets the problem of limited pre-training dataset and benchmarks.

In the healthcare domain, leveraging general domain pre-trained weights for downstream tasks is not straightforward for a low-resource language like Vietnamese. There are some attempts to populate Vietnamese healthcare corpora. The COVID-19 NER (COVID-19 Named Entity Recognition for Vietnamese) dataset (Truong et al., 2021) is the first reported dataset related to the healthcare domain, which is based on news text. Then the ViMQ (Vietnamese Medical Question) dataset (Huy et al., 2021) is introduced for developing healthcare chatbots, which is originated from medical questions. Anyway, NLP for the Vietnamese healthcare domain is still juvenile due to some challenges, including the lack of centralized storage for healthcare corpora like Pubmed[1]. Electronic Health Records (EHRs) are limited by outdated healthcare systems.

Due to the above concerns, we introduce acrDrAid, a human-labeled dataset for the Acronym Disambiguation (AD) task. This is the first AD dataset in Vietnamese, to the best of our knowledge. To examine the model in the text generation task, we propose Frequently Asked Questions (FAQ) Summarization data to summarize FAQ questions. We also empirically demonstrate the performance of the model on domain-specific downstream tasks with a multi-task learning strategy. We also show our experiments in extracting sentences from a large-scale corpus related to our target domain. The model and its variations are publicly released in `transformers`[2] (Wolf et al., 2020) un-

[1] `http://www.ncbi.nlm.nih.gov`
[2] `demdecuong/vihealthbert-base-word`

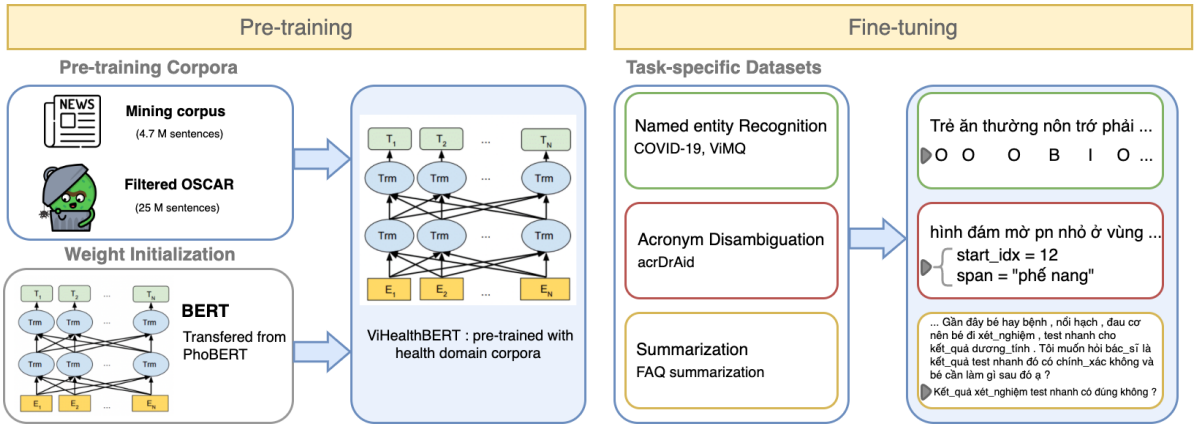


Figure 1: Overview of the pre-training and fine-tuning of ViHealthBERT.

der the name ViHealthBERT. The overall process of pre-training and fine-tuning ViHealthBERT is demonstrated in Figure 1.
Our contributions are summarized as follows:

- We present the first monolingual pre-trained Vietnamese language model for healthcare.
- We empirically investigate our model with different training strategies, achieving state-of-the-art (SOTA) performances on four downstream tasks: COVID-19, Medical Question Answering, Acronym Disambiguation, and Summarization.
- We introduce two Vietnamese datasets: the acrDrAid dataset and the FAQ summarization dataset in the healthcare domain. Our acrDrAid dataset is annotated with 135 sets of keywords.
- We experiment two approaches for the text mining process to select more data in low-resource language on a specific domain.

## 2. Dataset

We introduce two datasets to benchmark our language model in the health domain: acrDrAid and FAQ Summarization. An example of both datasets is illustrated in Figure 2.

### 2.1. acrDrAid

Acronym Disambiguation is the task of correctly identifying the expansion of an acronym in a given context. arcDraid is a Vietnamese dataset for AD that contains radiology reports from Vinmec hospital[3], Vietnam. Upon typing radiology reports, radiologists tend to use acronyms to save time. However, patients do not benefit from such acronyms due to the lack of medical background, leading to an incomplete understanding of the report. arcDrAid aims to facilitate the development of acronyms expansion tools, which helps both parties, radiologists and patients. Radiologists can speed up their typing process while patients still receive a full report with good transparency.

| Properties | Train | Dev | Test |
|---|---|---|---|
| no. samples | 4000 | 523 | 1130 |
| average input length | 30.52 | 29.92 | 32.23 |
| no. unique acronym | 109 | 109 | 135 |
| average no. expansion per acronym | 3.16 | 3.19 | 3.04 |
| no. of unique expansion | 276 | 182 | 279 |
| average acronym expansion length | 2.064 | 2.064 | 2.067 |
| prob. overlap acronym over train set | 100% | 100% | 80.74% |
| prob. overlap expansion over train set | 100% | 97.8% | 77.06% |

Table 1: Statistics of acrDrAid dataset. The stats are calculated in syllable-level unit.

Before annotation, a dictionary of acronyms and their possible expansions are generated as follows: All noun phrases (NP) from the radiology reports corpus are extracted with [4]underthesea pos-tagger. The 1000 candidates of 75000 NPs with the highest occurrence frequencies in the radiology reports corpus are selected for the dictionary keys using their abbreviations. Each dictionary entry refers to a list of possible expansions using such initials found in the radiology reports corpus. Entries with only one expansion are removed.
In the annotation process, we employ three expert radiologists to verify and resolve potential errors in the dictionary. Each sample contains an acronym and the paragraph in which it resides. The annotators have to verify the samples if it satisfies all three conditions:

[3] https://vinmec.com/

[4] https://github.com/undertheseanlp/underthesea: Open-source Vietnamese Natural Language Process Toolkit

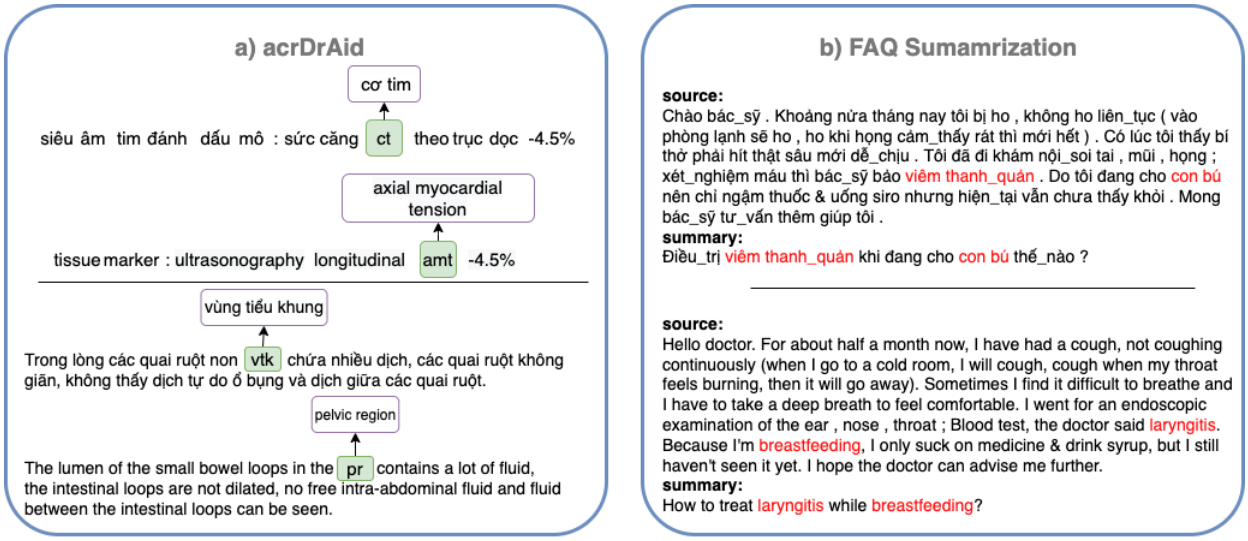


Figure 2: Illustration of our proposed datasets. a) An example of acrDrAid dataset. b) An example of FAQ dataset. The highlighted texts are words that are included in the source text. The English texts are the translation of the Vietnamese texts.

- The expansion for the acronym is correctly aligned with the sentence meaning.
- The expansions belong to the medical domain.
- The expansions are frequently used in radiology reports.

Each sample is verified by all three radiologists. We refer to a major voting scheme to finalize the label of the dataset.
To this end, we assume that replacing expansion words in a sentence with the corresponding acronym word will not change the sentence's meaning. We then randomly split the dataset into train/dev/test sets with a ratio of 7/1/2. The acrDrAid finally has 135 acronyms and 424 expansion texts in total. The overall statistic of the dataset is shown in Table 1.

## 2.2. FAQ Summarization

Text summarization is the task of shortening a long document while retaining concise, readability, and crucial information. In Frequently asked questions (FAQ) summarization, the summarizer not only selects an informative subset of words but also rephrases some typos and untechnical terms from the input question. This would reduce the workload of the doctor in categorizing a huge amount of questions and summarizing salient information. We introduce a sentence-level FAQ summarization dataset which is a semi manual-label dataset. This dataset is also a benchmark for the generation task of the models.
We first crawl articles in the FAQ section from healthcare trustworthy sites, including Vinmec[5] and Mevabe[6]. We use the titles as the summaries while the question text is the input sequences. The title is a summary of the content of the question, which makes the user easily get their needs. We then segment the crawled news articles using RDRSegmenter (Nguyen et al., 2018) from VNCoreNLP (Vu et al., 2018). The title of the question is human-written, so it satisfies the abstractive summarization property.
We keep the training set as its origin and normalize the gold label in the dev and test set. The normalizing process is to mitigate no bias in the gold label; we randomly get 1,500 samples to give to five annotators to label based on our guidelines. This process aims to refine spelling errors, missing punctuations, ambiguous summaries into a clearer version (e.g. before: "giá 5 in 1" ("price 5 in 1"); after "giá vaxcin 5 in 1" ("price of vaccine 5 in 1")) while retaining the natural cohesion of the summary and ensure the summary length must

| Properties | Train | Dev | Test |
|---|---|---|---|
| no. samples | 10621 | 1326 | 1330 |
| avg. input length † | 93.3 | 90.28 | 94.01 |
| avg. input length ‡ | 5.39 | 5.22 | 5.4 |
| avg. summary length † | 9.8 | 10.07 | 9.88 |
| avg. summary length ‡ | 1.02 | 1.02 | 1.04 |
| prob. target token is included in input | 77.7% | 78.17% | 71.81% |

Table 2: Statistics of FAQ Summarization dataset. prob. stands for probability. † and ‡ refers to word-level and sentence-level respectively.

[5] https://www.vinmec.com

[6] https://www.mevabe.vn

| Dataset | # Sent | Domain |
|---|---|---|
| Vietnamese Wikipedia | 5M | General |
| Vietnamese news | 96M | General |
| Our mining text (our) | 4.7M | Health, Medical |
| OSCAR's selected corpus (TF) | 25M | Health, Medical, General |

Table 3: Statistics of our pretraining dataset. M stands for million.

be 1-2 sentences with less than 15 words. To finalize the dataset, we have a meeting to discuss the annotated labels and finalize the summary. This makes the summary near human-like, which is one of the properties of abstractive summarization. The dataset has a total of 13277 samples, and the train/dev/test sets have a ratio of 8/1/1; the statistics of our dataset are presented in Table 2. Examples of acrDrAid and FAQ datasets are shown in Figure 2.

# 3. ViHealthBERT

This section describes the architecture, pretraining data, training strategy, and optimization setup for ViHealthBERT.

## 3.1. Architecture

In this work, our ViHealthBERT has only one version is ViHealthBERT$_{base}$ using the same architecture of BERT$_{base}$ (Devlin et al., 2019), which has 12 stacked encoder-only Transformer (Vaswani et al., 2017) layers, 768 hidden units and 12 attention heads. The large version will be updated in future works.

## 3.2. Pretraining data

### 3.2.1. Selected corpus

The selected pretraining dataset is a concatenation of two corpora : (i) our text mining corpus and (ii) selected corpus from parts of OSCAR dataset (Ortiz Suárez et al., 2019). Since we continually use the weight of PhoBERT for self-supervised training, the pretraining datasets of PhoBERT named Vietnamese Wikipedia and Vietnamese news[7] are also listed. The Vietnamese Wikipedia dataset is the Vietnamese edition from Wikipedia Corpus used by mBERT(Devlin et al., 2019) and XLM-R(Conneau et al., 2020). The Vietnamese news dataset is mined from many topics of news cites, but the healthcare-related topics are not included. The statistic of our pretraining dataset is presented in Table 3.

[7] https://github.com/binhvq/news-corpus crawled from a wide range of news websites and topics.

### 3.2.2. Text mining corpus

The corpus is gathered from a wide range of sources that are specific to the medical domain, including Vietnamese online news and hospital websites, scientific journals, and textbooks. For online news and hospital websites, articles from healthcare/medical sections are crawled using keywords such as health, vaccine, medical, and biomedical. For scientific journals, we extract the abstract of each article from Vietnamese Medical Journal[8], Journal of Health and Development Studies[9], Journal of Medicine and Pharmacy[10], and Medical Journal of Ho Chi Minh City[11].a For textbooks, we crawled the pdf version[12] and parsed them into structured text. This corpus is carefully pre-processed to clean noise, delexical emails, telephones, URLs, and deduplicate similar sentences via edit-distance. The cleaned corpus is used for self-supervised training of the model. The statistics of our mining data are shown in Table 3. More details on the sources and categories of the corpus are in Section Appendix.

### 3.2.3. Data Selection

This section describes how we selected healthcare-related sentences from the public dataset. Since data mining is a laborious process, we also want to mitigate that problem by utilizing a public dataset. We extract healthcare-related sentences from 32GB Vietnamese Corpus of OSCAR dataset (Ortiz Suárez et al., 2019) with two approaches: Term-Frequency (TF) method and Selector. The TF approach builds a high-quality dictionary from our mining corpus and then uses Term Frequency to select sentences based on the reference dictionary. The Selector is a classifier model that learns the joint domain of healthcare and general to classify a new sample. Almost all Vietnamese public datasets, such as (Raffel et al., 2020) and (Ortiz Suárez et al., 2019) are of the general domain, which is crawled from online news; selecting the target domain from a large corpus is also a challenge. We also make a comparison between dictionary-based(TF) and deep learning(Selector) on a manual-labeled set extracted from the OSCAR corpus, which is shown in Table 7.

**Term-Frequency (TF)**

In terms of TF, we hypothesize that high-quality keywords can ensure recall while mining a large corpus. Moreover, this approach is simple, low-cost, and fast, suitable for the industry. First, we build a dictionary from our mining corpus and select the top 3000 words that have over 10000 occurrences with the highest Term Frequency by manually observing the words related to the healthcare domain. Second, we manually filter these keywords to ensure it only remains words in the healthcare domain. These keywords also have been

[8] https://tonghoiyhoc.vn
[9] https://jhds.vn
[10] https://jmp.huemed-univ.edu.vn
[11] https://yhoctphcm.ump.edu.vn
[12] https://yhoctonghop.vn

double-checked by a medical graduate student. To this end, we assume that we have a high-quality dictionary of keywords for the healthcare domain for the term-frequency approach.

**Selector**

Since SimCSE(Gao et al., 2021) is a simple but effective method for learning sentence-level representation. We utilize this method to make our model aware of joint general-healthcare representation to classify healthcare-related sentences. This method has two variant approaches named Unsupervised and Supervised. The Unsupervised SimCSE approach uses different hidden dropout masks as noise for input sentences and makes the model predict itself from in-batch negatives. While the Supervised SimCSE leverages the NLI datasets and takes the entailment pairs as positives, contradiction pairs and other in-batch instances as negatives.

Regarding the Selector, we randomly select an equal size dataset from the Vietnamese news corpus (which has been used for training PhoBERT) and our mining corpus. We split our corpus and the selected set into *train-unsup* set for unsupervised training and *train-sup* set for supervised training. Then, we train an unsupervised encoder that follows the SimCSE(Gao et al., 2021) strategy with the *train-unsup* set. We utilize the encoder to map input into a joint embedding space and feed the output of the encoder into the classifier. We train a classifier that takes the output of the encoder on the *train-sup* set. The trained classifier is used to select the healthcare domain sentences from the large corpus.

### 3.3. Training Strategy

To enhance the domain-specific representation of the model, we train our model with multi-task learning for word-aware tasks are masked token (MLM) (Taylor, 1953) and capitalized prediction (CP). For masked tokens, the model has to reproduce the original masked token following the RoBERTa strategy (Liu et al., 2019). To learn a document structure-aware representation, we train the models with the next sentence prediction (NSP) task. We hypothesize that capital words usually have more semantic information, especially in specific domains (Sun et al., 2020). Predicting whether a word is capitalized or not can help the model become aware of the entities better.

We also experiment with two strategies for tokenizing the pretraining dataset: syllable-tokenizer and word-level tokenizer. A syllable tokenizer is a simple space-splitting tokenizer. For word-level tokenizer, we follow Nguyen and Tuan Nguyen (2020) and use RDRSegmenter as tokenizer. To the best of our knowledge, we are the first study that shows the performance of the Vietnamese monolingual pre-trained language model in two pre-processing approaches.

### 3.4. Optimization

We set the maximum length at 256 subtokens. We optimize the models using AdamW(Loshchilov and Hutter, 2019) and a learning rate of 1e-5. We initialize our model with PhoBERT(Nguyen and Tuan Nguyen, 2020) and a batch size of 64 for all settings across one server A100-40GB. The models are trained with 400k steps for three days on the mining corpus For the combination of the mining corpus and TF corpus, we trained the model for one epoch in six days.

Since there are multi-task training settings, we set the learning of both MLM and CP are fixed as 0.15. For setting has two pretraining tasks, the learning rate for MLM is 0.85. For the experiment containing three pretraining tasks, the MLM learning rate is set at 0.7.

## 4. Experimental Setup

We evaluate the performance of ViHealthBERT on three tasks using four downstream Vietnamese corpora: COVID-19, ViMQ, acrDrAid, and FAQ Text Summarization.

### 4.1. Downstream task datasets

The COVID-19 NER dataset(Truong et al., 2021) is the first manually annotated domain-specific dataset on Vietnamese. This dataset contains 10 entity types with a total of 10027 samples that are divided into train/valid/test samples with the amount of 5027/200/3000, respectively. We use this dataset in word-level version as same as PhoBERT Nguyen and Tuan Nguyen (2020).

In terms of ViMQ(Huy et al., 2021), this dataset provides both intents and named entities labels for Vietnamese medical questions. In this work, we evaluate our model only on the NER task. This dataset consists of 3 entity categories: symptom-disease, medicine, and medical procedure. The dataset is split into train/dev/test with the amount of 7000/1000/1000 and only has the word-level version.

For the acrDrAid dataset, this is an in-house dataset consisting of 135 acronyms and 424 expansions. This dataset is divided into 4000 train samples, 523 validation samples, and 1130 test samples. The average number of expansions per acronym of the training and test sets is 3.16 and 3.04. This dataset is syllable-tokenized and has the probability of overlapping acronyms between the test set and the train set is 80.74%.

For the FAQ Summarization, we employ RDRSegmenter(Nguyen et al., 2018) to segment the text into words before applying the BPE tokenizer. The dataset contains human questions from faq sections on online sites. It is split into 10621 train samples and 2656 test samples. On average, there are approximately 93.3 syllables per sample and 9.8 syllables per reference summaries in the training set. The average number of sentences per input document of the training and test sets are 5.39 and 5.31, respectively.

### 4.2. Fine tuning

Following Truong et al. (2021) for NER, we finetune our models with a fixed learning rate of 5e-5 and a

batch size of 32 for 30 epochs. We evaluate the task performance after each epoch on the validation set with metric-triggered early stopping after 5 epochs. The best checkpoint is then evaluated on the test set to report the final score. The reported score is the average score of each best score from 5 finetune times with different random seeds.

In the ViMQ dataset, we finetune our models with a fixed learning rate of 5e-5 and a batch size of 32 for 10 epochs. We evaluate the task performance after each epoch on the validation set with early stopping after 5 epochs based on the highest F1-score. The best checkpoint is then evaluated on the test set to report the final score. We get the average score of each best score from 3 different random seeds fine-tuning turns.

In the acrDrAid benchmark, we finetune the models on 10 training epochs with a fixed learning rate of 1e-5 and batch size of 32. We select the best checkpoint on the validation set after each epoch and evaluate it on the test set. The reported score is averaged after 3 times running with different random seeds.

In FAQ Summarization, we use our models as the encoder. We use 12 stacked random initialize Transformer layers for the decoder, and the beam size is 5. The models are trained with a batch size of 64, and a learning rate of 1e-5. We finetune in 10 training epochs then get the checkpoint that has the highest ROUGE-average(Lin, 2004) score on the validation set and evaluates it on the test set. The reported result is the average performance after initializing with 3 different random seeds.

For COVID-19, ViMQ and acrDrAid, the assessment settings are conducted across single sever A100-40GB. For FAQ Summarization, all the fine-tuning experiments used P100-16GB.

### 4.3. Data Selection

We split our corpus and the selected set from Vietnamese corpus into *train-unsup* set and *train-sup* set with ratio 9/1. The classifier module is simply a Multiple Layer Perceptron on top of the trained encoder. The encoder follows RoBERTa architecture that has been initialized randomly. First, we train the encoder following unsupervised strategies as same as SimCSE (Gao et al., 2021). Then for each epoch, we train the encoder and the classifier with *train-sup* dataset followed by k-fold cross-validation. We set k=5, and the train/dev/test ratio is 4/1/5. The last checkpoint of each epoch with the highest average score over k-fold is saved and used to classify sentences from the unseen corpus. In this work, we experiment only with a corpus that is selected from TF, and we would discuss the reason in Section 5.2.3.

## 5. Experimental Results

### 5.1. Main result

In terms of the COVID-19 dataset, our models outperform the original pre-trained PhoBERT in all scores across all settings. Our pre-trained models with word-level datasets improved about 2.2% in Mac-F1 and 4.7% in Mic-F1. We observe that the models show no significant improvement within different pre-training tasks. Besides, training with our mining corpus and TF corpus does not hurt the model performance, which means the dictionary-based method for selecting the target domain in unseen data is a potential approach.

For the ViMQ dataset, almost all scores of our models exceed PhoBERT. Compared to models trained with word-tokenizer, models trained with syllable-token show better performance in almost all settings. In terms of the highest model, the model shows the competitive result by the higher score in Mac-F1 and Mic-F1 of +2.26% and +3.97%, respectively. The overall result is shown in Table 4.

For the acrDrAid dataset, we observe that all the models have higher scores in all metrics compared to the baseline, which is presented in Table 5. ViHealthBERT helps boost the performance of the PhoBERT$_{base}$ across Mac-Pre, Mac-Rec, and Mac-F1 by +1.17%, +6.28%, and +4.19%, respectively. Compared to PhoBERT$_{large}$, our models also show better performance while enhancing the Mac-Pre, Mac-Rec, and Mac-F1 by +1.64%, +10.71%, and +7.14%, respectively.

For the FAQ summarization dataset, all settings of ViHealthBERT show consistent improvement. In terms of the best setting that has been trained with MLM, the model shows the increase in R1, R2, R-L are +3.3%, +3.17%, and 2.66%, respectively. The overall result is shown in Table 6

### 5.2. Discussion

#### 5.2.1. Multi-task learning

According to Table 4, we find that multi-task learning does not boost up the performance across NER tasks. For the AD task, we observe that NSP shows better results in macro-Recall and macro-F1 while achieving comparable Precision. For FAQ summarization, by enhancing entity aware and structure-aware ability of the model through CP and NSP tasks, we achieve a comparable performance compared to the model that used MLM tasks only. In general, the experiment results demonstrate the modest benefits of multi-task learning in certain tasks, but not all of them.

#### 5.2.2. Tokenizer for pretrained data

We observe that the tokenization level of the pre-training data can affect the performance of the model in downstream tasks. In ViMQ and acrDrAid datasets, the models that use syllable-level tokenizer are better than those that use word-level tokenizer for all settings. In the case of COVID-19, there is no significant difference among models beyond the two types of tokenizers, possibly indicating a saturated performance of the model on this dataset. Our experiments show that only using syllable tokenizing could lead to beneficial gain

| Model | Tokenize-level | Pre-training data | Pre-training task | COVID-19 | | ViMQ | |
|---|---|---|---|---|---|---|---|
| | | | | Mac-F1 | Mic-F1 | Mac-F1 | Mic-F1 |
| PhoBERT$_{base}$ | word | * | MLM | 0.942 | 0.920 | 0.8470 | 0.8224 |
| PhoBERT$_{large}$ | word | * | MLM | 0.945 | 0.931 | 0.8524 | 0.8257 |
| ViHealthBERT | word | * + our | MLM | **0.9677** | **0.9677** | 0.8601 | 0.8432 |
| ViHealthBERT | word | * + our | MLM + NSP | 0.9674 | 0.9674 | 0.8562 | 0.8441 |
| ViHealthBERT | word | * + our | MLM + CP | 0.9677 | 0.9677 | 0.8578 | 0.8397 |
| ViHealthBERT | word | * + our | MLM + NSP + CP | 0.9662 | 0.9619 | 0.8526 | 0.8383 |
| ViHealthBERT | syllable | * + our | MLM | 0.9652 | 0.9653 | 0.8575 | 0.8481 |
| ViHealthBERT | syllable | * + our | MLM + NSP | 0.9673 | 0.9673 | 0.8610 | 0.8440 |
| ViHealthBERT | syllable | * + our | MLM + CP | 0.9672 | 0.9677 | 0.8664 | 0.8567 |
| ViHealthBERT | syllable | * + our | MLM + NSP + CP | 0.9665 | 0.9664 | 0.8676 | 0.8641 |
| ViHealthBERT | syllable | * + our + TF | MLM | 0.9639 | 0.9641 | **0.8698** | **0.8621** |
| ViHealthBERT | syllable | * + our + TF | MLM + CP | 0.9629 | 0.9629 | 0.8632 | 0.8501 |

Table 4: The overview of experimental results in COVID-19 and ViMQ datasets. * refers to pretrained dataset of PhoBERT (Nguyen and Tuan Nguyen, 2020).

| Model | Tokenize-level | Pre-training data | Pre-training task | Mac-Pre | Mac-Rec | Mac-F1 |
|---|---|---|---|---|---|---|
| PhoBERT$_{base}$ | word | * | MLM | 0.9197 | 0.7481 | 0.8251 |
| PhoBERT$_{large}$ | word | * | MLM | 0.9150 | 0.7038 | 0.7956 |
| ViHealthBERT | word | * + our | MLM | 0.9281 | 0.7565 | 0.8336 |
| ViHealthBERT | word | * + our | MLM + NSP | 0.9320 | 0.7562 | 0.8349 |
| ViHealthBERT | word | * + our | MLM + CP | 0.9235 | 0.7605 | 0.8341 |
| ViHealthBERT | word | * + our | MLM + NSP + CP | 0.9212 | 0.7583 | 0.8318 |
| ViHealthBERT | syllable | * + our | MLM | 0.9291 | 0.7928 | 0.8555 |
| ViHealthBERT | syllable | * + our | MLM + NSP | 0.9314 | **0.8109** | **0.8670** |
| ViHealthBERT | syllable | * + our | MLM + CP | **0.9402** | 0.7838 | 0.8559 |
| ViHealthBERT | syllable | * + our | MLM + NSP + CP | 0.9362 | 0.777 | 0.8493 |
| ViHealthBERT | syllable | * + our + TF | MLM | 0.9315 | 0.7878 | 0.8537 |
| ViHealthBERT | syllable | * + our + TF | MLM + CP | 0.9287 | 0.7927 | 0.8553 |

Table 5: The overview of experimental results in acrDrAid datasets. * refers to pretrained dataset of PhoBERT (Nguyen and Tuan Nguyen, 2020).

for downstream tasks. For a medium-sized, clean corpus, a word-tokenizer is an optimal solution because a word is the core semantic unit in Vietnamese. For large corpus with noisy unstructured and informal text, a syllable tokenizer is a more suitable choice due to typos, misspellings. Tokenization is an important factor when fine-tuning pre-trained models.

#### 5.2.3. Data selection

In this section, we will explain the reason for choosing the corpus selected by TF only. We compare two methods, TF and Selector, on the OSCAR dataset by creating a manual test set. We use both mentioned methods to inference a subset of data before randomly collecting 3000 samples from each method for examination. In the examination phase, we have strictly defined guidelines to distinguish sentences that belong to the health domain or general domain. According to Table 7, the Selector does not become aware of the health domain samples well in unseen data while TF approach has selected 83.5% healthcare-related sentences over 2276 sentences[13].We hypothesize that the distribution of unseen data is too diverse and chaos that makes the model not generalize well. The result of Selector shows that active learning on a large corpus is one of the challenges for a specific-domain language model.

We find that with more data selected from the OSCAR corpus, the performance of the model shows no significant changes. The models trained with the combination corpus of our text mining and TF show comparable performance to those trained with our mining corpus. We hypothesize that our mining corpus's quality and diversity are still suitable for our benchmark dataset.

## 6. Conclusion

We introduced, ViHealthBERT, a pre-trained language representation model in Vietnamese health news text mining. We show that pre-training BERT on health corpora is crucial in applying it to the healthcare

[13]We also use this manual labeled data to train supervised the Selector but the performance do not show improvement.

| Model | Tokenize-level | Pre-training data | Pre-training task | R-1 | R-2 | R-L |
|---|---|---|---|---|---|---|
| PhoBERT$_{base}$ | word | * | MLM | 47.15 | 28.18 | 41.16 |
| ViHealthBERT | word | * + our | MLM | **50.45** | **31.35** | **43.85** |
| ViHealthBERT | word | * + our | MLM + NSP | 49.06 | 29.7 | 42.56 |
| ViHealthBERT | word | * + our | MLM + CP | 47.85 | 29.81 | 42.04 |
| ViHealthBERT | word | * + our | MLM + NSP + CP | 50.4 | 31.15 | 43.78 |
| ViHealthBERT | syllable | * + our | MLM | 48.47 | 28.65 | 41.77 |
| ViHealthBERT | syllable | * + our | MLM + NSP | 49.92 | 30.77 | 43.37 |
| ViHealthBERT | syllable | * + our | MLM + CP | 48.56 | 29.20 | 41.96 |
| ViHealthBERT | syllable | * + our | MLM + NSP + CP | 48.33 | 30.35 | 42.36 |
| ViHealthBERT | syllable | * + our + TF | MLM | 48.32 | 29.17 | 42.07 |
| ViHealthBERT | syllable | * + our + TF | MLM + CP | 49.45 | 30.86 | 43.26 |

Table 6: The overview of experimental results in FAQ datasets. R-1, R-2, R-L refers to ROUGE-1, ROUGE-2 and ROUGE-L respectively. * refers to pretrained dataset of PhoBERT(Nguyen and Tuan Nguyen, 2020).

| | **# Samples** | **Selector** | **TF** |
|---|---|---|---|
| General | 3724 | 2625 | 1099 |
| Health | 2276 | 375 | 1901 |
| Total | 6000 | 3000 | 3000 |
| Prob. health | 37.9% | 12.5% | 63.37% |

Table 7: The overview result of data selection. Prob. health refers probability of health samples over total samples.

and biomedical domain. Our ViHealthBERT demonstrates potential results by producing the SOTA performance in four health-domain Vietnamese datasets: COVID-19, ViMQ, acrDrAid, and FAQ Summarization. We also experience two approaches to select specific-domain in a large public corpus are TF and Selector. We release two monolingual datasets for future researches of NLP in the health domain: acrDrAid and FAQ summarization. We hope that ViHealthBERT will be a strong baseline for future Vietnamese NLP research and applications in the healthcare or medical domain.

## Appendix

### Details on our mining corpus

In this section, we describe our mining corpus in detail. The overall of categories is shown in Table 8 and the common mining cites are presented at Figure 3. We focused on five categories: online news, hospital cites, specialized news, journals, and textbooks. For each category, the following definition is:

- Online News: Digital documents from legit Vietnamese publishers (e.g. vnexpress.net, tuoitre.vn)
- Hospital Cites: Online main sites of Vietnamese hospital (e.g. benhvien115.com.vn, vinmec.com)
- Specialized News: "expert" cites in the biomedical field, almost all of theses pages are responsible for a dictionary of disease, medicine (e.g. ykhoa.net)
- Journals: Science publications or research are written in Vietnamese (e.g. jhds.vn, vnpharmjour.org.vn)
- Textbooks: Crawled specialized textbooks(from yhoctonghop.vn). Due to their are pdf, we have to parse them into text and structurize them before put into the pre-training corpus.

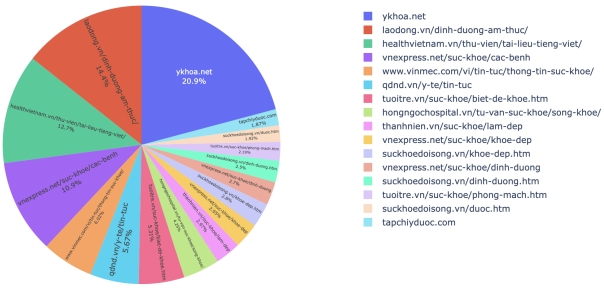


Figure 3: The proportion of most common domains cites across 89.2% our mining corpus. These cite quantities make up about 28.6% of the total mining cite.

## 7. Bibliographical References

Bui, T. V., Tran, T. O., and Le-Hong, P. (2020). Improving sequence tagging for vietnamese text using transformer-based neural models. In Minh Le Nguyen, et al., editors, *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation, PACLIC 2020, Hanoi, Vietnam, October 24-26, 2020*, pages 13–20. Association for Computational Linguistics.

Chau, C.-N., Nguyen, T.-S., and Nguyen, L.-M. (2020). Vnlawbert: A vietnamese legal answer selection approach using bert language model. In

| Category | Domain | Linguistic Style | Size(MB) | Tags |
|---|---|---|---|---|
| News | Health, Medical, General | Generic | 471.66 | beauty, health, nutrition, diseases, vaccines, clean eating, health living, sex, news |
| Hospital Cites | Medical, Health | Generic | 165.7 | vaccines, diseases, dictionaries, customer stories |
| Specified News | Health, Medical | Generic, Academic | 184.8 | medicine, beauty, nutrition, disease, gender, dictionary, medical knowledge |
| Journals | Medical | Academic | 7.8 | abstract (problem, objective, object, method, result, conclusion, keyword) |
| Text Books | Medical | Academic | 9.8 | full corpus |

Table 8: Our mining corpus in detail. Tags are section name that we concentrate on mining.

*2020 7th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 298–301.

Clark, K., Luong, M., Le, Q. V., and Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Gao, T., Yao, X., and Chen, D. (2021). SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. cite arxiv:1907.11692.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nguyen, D. Q. and Tuan Nguyen, A. (2020). PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online, November. Association for Computational Linguistics.

Nguyen, D. Q., Nguyen, D. Q., Vu, T., Dras, M., and Johnson, M. (2018). A Fast and Accurate Vietnamese Word Segmenter. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 2582–2587.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Roy, A. and Pan, S. (2021). Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., and Wang, H. (2020). ERNIE 2.0: A continual pre-training framework for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8968–8975. AAAI Press.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July. Association for Computational Linguistics.

## 8. Language Resource References

Huy, T. D., Tu, N. A., Vu, T. H., Minh, N. P., Phan, N., Bui, T. H., and Truong, S. Q. H. (2021). Vimq: A vietnamese medical question dataset for healthcare dialogue system development. In Teddy Mantoro, et al., editors, *Neural Information Processing*, pages 657–664, Cham. Springer International Publishing.

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Truong, T. H., Dao, M. H., and Nguyen, D. Q. (2021). COVID-19 named entity recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2146–2153, Online, June. Association for Computational Linguistics.

Vu, T., Nguyen, D. Q., Nguyen, D. Q., Dras, M., and Johnson, M. (2018). VnCoreNLP: A Vietnamese Natural Language Processing Toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.