Received: June 12, 2023. Revised: August 23, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Comparison of Transformer and BiLSTM Based BioNER Model with Self-Training on Low-Resource Language Texts of Online Health Consultations

Diana Purwitasari[1, 2]* Abid Famasya Abdillah[1] Safitri Juanita[3,4]
I Ketut Eddy Purnama[2,5] Mauridhi Hery Purnomo[2,3,5]

[1]*Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[2]*University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS), Indonesia*
[3]*Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
[4]*Department of Information System,Universitas Budi Luhur, Jakarta, Indonesia*
[5]*Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
* Corresponding author's Email: diana@if.its.ac.id

Abstract: More people have gotten used to online health consultations (OHC) because of the COVID-19 pandemic to reassure themselves of their health conditions or seek other treatment options. The OHC system could use named entity recognition (NER), specifically for health-related texts called biomedical NER (BioNER), to filter text entities from posting history to ease users' finding information. The terms of named entities (NEs) could be related to human anatomy that have some inconvenience or terms to find out any symptoms of the disease. However, OHC posts, especially user questions, are often non-formal sentences and even long sentences or have incorrect medical terms since the users are most likely non-trained medical professionals, which may lead to out-of-vocabulary (OOV) problems. Although long short-term memory (LSTM) architecture is known for its advantage in modeling sequential data like text, and even with the bidirectional version of BiLSTM, it has some difficulties handling those long sentences. A transformer model could overcome the problems. Another problem concerns fewer annotated data in low-resource language OHC texts despite data abundance in the corpus crawling from the OHC platform. To augment data training, our process includes a self-training approach as semi-supervised learning in data preparation to improve a BioNER model. In preparation for our BioNER model, this work observes and makes a comparison on the embedding step, whether stacked embedding of BiLSTM-based or fine-tuning of transformer-based and defines filtering pseudo-labels to reduce noise from self-training. Although the empirical experiments utilized OHC texts in Indonesian as a case of low-resource language texts because of our familiarity, the procedures in this work apply to Latin alphabet-based languages. We also observed other biomedical NER model creation and topic modelling for verifying the extracted entities from the resulted BioNER model to validate the procedures. The results indicate that our framework, which includes preparing data from raw texts into labelled data using self-training, with a confidence threshold of 0.85, to create the BioNER model, can give F1 scores of 0.732 and 0.838 for BiLSTM-based and transformer-based models.

Keywords: Online health consultation texts, Named entity recognition, Low resource language, Semi-supervised learning.

1. Introduction

As a result of the COVID-19 pandemic, many individuals have become accustomed to online health consultations (OHC). The government can monitor frequently asked questions in online consultation platforms and other publicly available health information channels to gain insight into public concerns. For instance, a system called SENTINEL can detect outbreaks by analyzing Twitter data [1]. Similarly, some words become Google Trends because they are often used as search keywords when certain events happen. With the observed application, those events were analyzed to have topical relations to the trending words [2]. Thus, the topic keywords extraction gathers and analyses societal health

condition cues. However, the topic keywords taken from OHC texts could serve a different purpose. Topic keywords or representative words in topic modelling will become the topic label and assist information seekers in finding relevant responses from doctors or physicians in the OHC platforms, enabling them to comprehend their health status better [3]. Through OHC postings, doctor interactions become social support for the users or patients looking for other treatment options [4]. These studies define the support keywords based on specific groups, such as emotional or informational support, service attitude, or other satisfaction-related variables.

Besides those support-defined categories, the topic keywords in OHC texts could comply with a standardized categorization of unified medical language system (UMLS) guidelines, such as entities of symptoms, body parts, or chemicals [5]. For entity recognition, the named entity recognition (NER) model extracts general-domain entities and BioNER (biomedical named entity extraction) extracts domain-specific biomedical entities from unstructured health or medical-related texts. NER is a core component of a question-answer system since it extracts entities to support questions with simple facts, i.e., location, person, or organization [6]. For OHC texts, BioNER is more fitting. The OHC system could provide answer suggestions based on history and give relevant discussions to the information seekers by filtering any existing entities of the question using BioNER. All mentions related to abnormalities (i.e., symptoms, dysfunction, disease) are grouped into the DISO tags, and human anatomy-related mentions (i.e., body part, cell, tissue) are grouped into the ANAT tags.

Recent studies on BioNER showed a deep learning method with a collaborative strategy on a high-resource language of Chinese on OHC texts [7]. The studies have addressed the large communication volume between patients and doctors by creating virtual health assistants using previous consultation texts to ease the burden of doctor workloads and increase medical efficiency. To overcome out-of-vocabulary (OOV) problems (i.e., “headache” and “dizzy” that could be treated differently), the classic combination of word-embedding and long short-term memory (LTSM) deep learning were combined or improving transformer-based deep learning that could handle ambiguity caused by nonstandard medical entities. Those studies of high-resource language texts demonstrated that data availability and Chinese language pre-trained data make entity recognition succeeds. Medical experts performed annotation steps in those BioNER studies of high-resource language texts like Chinese [8] or English [9, 10]. The annotation strategy is necessary in the initial stage and even more for NER on texts with low-resource languages like Arabic, Vietnamese, Polish, Japanese, and Indonesian [11].

Research on extracting clinical entities such as biomedical for applications from another domain like OHC based on UMLS still has broad opportunities, especially for low-resource language. Studies of extracting Korean clinical entities from OHC texts have used bidirectional encoder representations for transformers (BERT) modified on postpositions to handle the Korean word structure [12]. The studies were on many languages and used bidirectional LSTM (Bi-LSTM) - conditional random fields (CRF) to include relations between sub-words [13] [14]. The Korean BioNER dataset in this work has been set up according to specialists categorized into medical departments. The specialists did manual annotating since this step is crucial for low-resource languages. An additional step is required to evaluate the different possibilities of annotations, i.e., works on Spanish texts of medicinal product characteristics with different entities like excipient, medicament, or therapeutic action, among others, have defined its evaluation procedures when testing the reliability of annotators [15]. Manual annotation needs human resources such that semi-supervised self-training augments the number of training data and enhances the performance of the NER model even with limited training data [16]. Supervised learning requires careful examination of the ruleset to generate labelled data, whilst semi-supervised will reuse features based on previously established supervised classifiers to make less noise apparent in the training data.

Motivated by the previous works on low-resource language and the need to utilize existing information that has accumulated e-health diagnoses from past cases answered by registered doctors in the OHC platform, we define a process for the BioNER model. The need for BioNER would adapt to the following situation. Each doctor could have different expression styles when giving advice, even for the same cases. Furthermore, the users of non-trained medical professionals might tell varied symptoms which lead to the same disease even for different categories. A doctor could answer repeating questions. Suppose the BioNER model has tagged biomedical entities from existing answers. In that case, the doctor may look up similar symptoms to shorten the searching time, point out related previous answers, and provide more detailed answers if required. Previous works have accentuated deep learning methods, especially on standard LSTM for text analysis and transformer-based to tackle

ambiguity or OOV problems [7, 8]. The LSTM approach is suitable for OHC texts by analyzing sequential data on non-formal sentences, especially with the bi-directional approach (BiLSTM). With various users and their differences in writing styles, transformer-based architecture such as BERT is preferable for the OOV problem.

The works of [2-4] exposed that keywords or topics as groups of keywords could help users navigate the contents of OHC texts. And it will be more beneficial if the keywords have been annotated, as shown in [5, 6]. Previous works of other low-resource languages [7-14] had more manually labelled sentences. With the technique of self-training [15, 16], we propose some steps in this work.

The contribution of this work is to investigate the effectiveness of known transformed-based compared to BiLSTM-based in a low-resource language, which in our case is Indonesian texts, inspired by previous works [12] and provide a BioNER dataset based on the OHC domain. We incorporate self-training to augment the dataset because of the limited size of the unannotated data. Self-training generates pseudo-labelled data to be added for data training, then re-train iteratively to a mix of initial data training and pseudo-labelled data. Since this strategy could introduce noise into the dataset, our evaluation procedure filters the entities by dropping pseudo labels lower than a threshold to reduce noise. Thus, this study aims to investigate the performance of an existing combination of steps without much manual preparation from raw OHC texts of low-resource language and depends on self-training to enrich the generated BioNER model using BiLSTM and transformer-based.

This section (*1. Introduction*) has discussed some steps required to set up a low-resource language BioNER model. The following section (*2. Related Works*) describes some other works about BioNER, even if they are outside OHC texts. We also mention a few of the existing BioNER models for comparison in our experiments. Then, we describe our processes (*3. Method*). The methods start from embedding, classifying with BiLSTM-based and transformed-based, and then adding data with a self-training approach using our thresholding criteria to get a refined BioNER model. Lastly, to verify the effectiveness of our model, we also observe it with other biomedical NER modeling methods (*4. Results and Discussions*). The observation still uses our data of low-resource language texts, which are OHC texts in Indonesian. Then we verify our generated data with our proposed framework using CrossNER [8] and CollaboNET [9]. We showed that the BioNER results of those low-resource data are comparable with the results of other standard datasets of high-resource language in [8, 9].

2. Related works

A NER tagger tool, HunFlair, uses a recent natural language processing (NLP) framework called flair with its pre-trained model similar to Bi-LSTM [10]. HunFlair identified entities of cell lines, chemicals, diseases, genes, and species, which were trained on a corpus of biomedical abstracts and full texts. For each entity type, associated corpora are required to enhance training. Flair has features like word embedding to character level, which could overcome the OOV problem, defined as Flair embedding, to perform NLP tasks.

Giving semantic context to each word could result in better vector representation. Therefore, word embedding with state-of-the-art Word2Vec is famous for its enhancement, like Fast Text, which works on sub-word embedding by splitting the original words. Byte pair encoding (BPE) approach defines the sub-word concept by merging recurring characters to create new symbols, i.e., Korean or Chinese characters [18]. The studies also have resulted a pre-trained model called BPEmb from training Wikipedia texts and supports hundreds of languages, including low-resource languages.

Fast Text and BPE find the closeness between the existing sub-word embedding and characters in new tokens that have yet to be learned, even though sometimes they are unknown terms. Accordingly, we utilize it in the stacked embedding or concatenate vectors from those pre-trained models to get better vector representation, and then then compare them with another character level embedding of Flair. The works of [10, 18] showed that Flair and BPEmb as embedding layer is suggested for the OOV problem and will be investigated within our framework.

The OOV problem could be caused by a phrase of words, such that some words will create a different word phrase and have different entity types depending on a particular context, i.e., a word phrase of "New York Times" for ORG (organization) and its sub-phrase of "New York" for LOC (location). conditional random fields (CRF) formulate structural dependencies between words as graphical models, and the solution employs the transition probabilities. As mentioned before, symbol-based words like Korean [13, 14] or Chinese [19] characters get the benefits of CRF. Each character depicts certain meaning. The latter studies combined transformer and LSTM based on the BioNER model. Motivated by those works, our empirical experiments observe the CRF usage in non-symbol-based words (A-Z) for

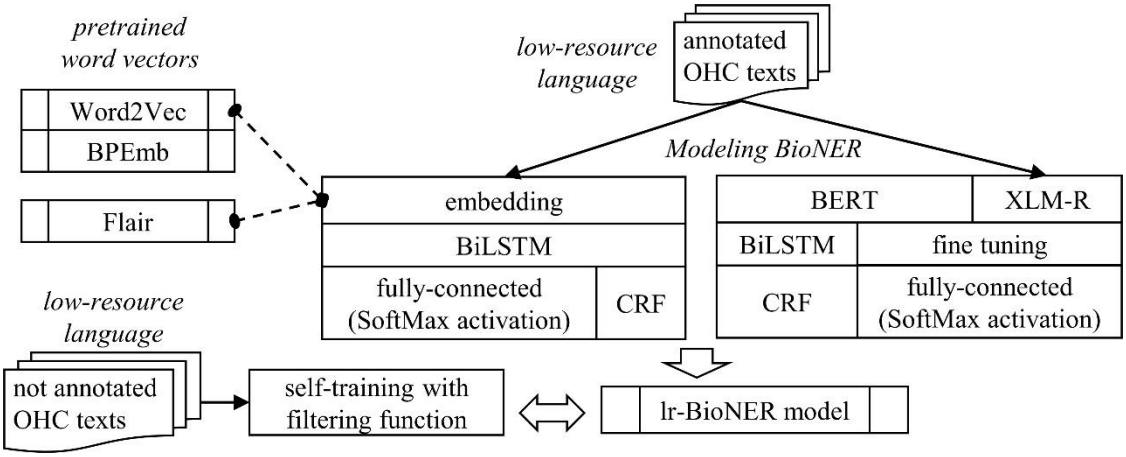


Figure. 1 Proposed framework for BioNER modeling with self-training on low-resource language texts of online health consultations (lr-BioNER)

BioNER model.

The works of [13-15] showed that for low-resource language, CRF as the final layer of Bi-LSTM or transformed could be explored and will be investigated within our framework as well.

Those previous works have trained the BioNER model on academic texts which tend to have formal sentences. In contrast, our work exhibits a dataset of OHC texts that are often non-formal and could be long sentences or have incorrect medical terms. Like BPE trained on the Wikipedia dataset, an extension of BERT called XLM-R (extended language modeling RoBERTa or robustly optimized BERT pretraining approach) [20] is a pre-trained model with a transformer-based multilingual on the CommonCrawl[1] dataset as high-resource language.

The OOV problem related to semantics is a sequence-to-sequence approach to capture words' context. Transformers consist of encoders-decoders blocks and could solve the OOV problem. The encoder applies self-attention to inputs to encode them into contextualized weight matrices. The main advantage of the encoder-decoder is that the pre-trained weights in each block can be fine-tuned to solve specific tasks, i.e., encoder for text summary, document classification, or NER tasks, and decoder for text generation and augmentation.

Despite the already trained model in some high-resource languages, we employ XLM-R with low-resource languages. The pre-trained model is fine-tuned with OHC data of low-resource languages to recognize specific UMLS entities (transfer learning) by using weights obtained from previous training steps to make developing a BioNER model more efficient. We observe the previous methods as a comparison in the empirical experiments, including XLM-R. We also observe other biomedical NER models based on CrossNER [8] and CollaboNet [9] for the empirical experiments. Both models utilize BiLSTM+CRF and add character-level embedding with a deep learning convolutional neural network (CNN) architecture. However, CrossNER modified the BiLSTM and made it into a hybrid of transformer and LSTM-based model. For CollaboNet, the NER task is like an ensemble model, with each named entity having a NER model.

3. Method

3.1 Data collecting

OHC texts contain pairs of questions by the public users and answers provided by authorized doctors on the OHC platform. There is no following question for each doctor in the platform, so each person could ask more than one question, but all questions are considered one long question text. The framework begins with a data-crawling from the OHC platform of low-resource language texts to build a base dataset, followed by a pre-processing pipeline. Although standard text pre-processing (i.e., stemming, lemmatization, and token removal) could reduce feature space complexity, the data cleaning process after crawling from the OHC platform for the NER model is conducted on the lengthy sentences of doctors' answers solely by removing non-alphanumeric characters and HTML tags, then trimming excessive use of space.

Data annotation is prepared by experts afterward to label a small amount of data. Inside-outside-beginning (IOB) format was used to follow standard annotation tags in NER tasks. This format also has the advantage of carrying out complex entity grouping over multiple word spans by placing a

[1] https://commoncrawl.org/

prefix tag boundary. For example, the sentence "Knee injury may happen to my left leg" will be tokenized and tagged with {B-DISO; I-DISO; O; O; O; B-ANAT; I-ANAT; I-ANAT}.

Terms related to symptoms or disease are grouped into the DISO category, like "knee injury". Then, human anatomy-related mentions are grouped into the ANAT tags, like "my left leg".

Question-answer (QA) texts in a OHC platform have topic categories t_c such as baby, pregnancy, skincare, or other topics. The entities e_i are extracted on the doctor's answers d_j since the BioNER task is a preliminary step to extend the OHC application, such as filtering QA texts to assist information seekers based on topics in specific categories. For training data, answer texts of OHC are preferable because the doctors' responses are more extended and have more expressions than the questions. Some single word or phrase words w_k in d_j (called as w_{jk}) can be labelled as entities, such as DISO or ANAT (i.e. each tuple word-entity tag is $\langle w_{jk}, tag_{jk} \rangle$ with $\langle$knee, B-DISO$\rangle$ and $tag_{jk} \in e_1 \ldots e_5$ such that all possibilities in here, e_1=O, e_2=B-DISO, e_3=I-DISO, e_4=B-ANAT, and e_5=I-ANAT).

3.2 Base BioNER modeling

Here, BioNER task is an information extraction to identify targets of user-specified entities by locating named entities in the unstructured OHC texts and classifying them into predefined labels of UMLS entities. Our work utilizes existing NLP frameworks to avoid task-specific solutions and emphasize the practical feasibility of the BioNER model deployed in an OHC platform (Fig. 1).

One crucial factor contributing to the NER task is extracting meaningful terms from source texts through word embeddings, which could be categorized as global and contextual. The global embedding approach catches syntactic and semantic latent similarities based on the co-occurrences of terms, such as the classic Word2Vec and Fast Text, that incorporate sub-word embedding for the OOV problem. The contextual embedding captures the uses of the same exact words in different contexts, like transformer-based pre-train embedding models.

Our BioNER model employs deep learning architectures of BiLSTM-based and transformer-based. Fig. 1 presents stacked embedding is used for feature extraction of BiLSTM-based (global embedding) and fine-tuning of transformer-based (contextual embedding). The BiLSTM architecture contains an embedding layer (Word2Vec, BPEmb, and Flair) that exploits pre-trained word vectors to transform inputs into dense vector representations of contextualized embedding weight based on its surrounding word. Thus, the combinations are Word2Vec-BPEmb and Flair. Generally, a feature matrix with rows of doctor's answers d_j and columns of indexed words w_k, the value of ω_{jk} is a weight for a certain word in a document w_{jk}. Classic word embedding makes ω_{jk} as a combination of term and document frequency. By using stacked embedding of pre-trained models, Fig. 1 shows Word2Vec and BPEmb, the weight value of each word is averaging from a concatenation of word vectors from the pre-trained models (1), i.e., $\underset{\omega 1_k}{\longrightarrow}$ from Word2Vec and $\underset{\omega 2_k}{\longrightarrow}$ from BPEmb.

$$d_j = [\omega_{jk}] = \left[\text{avg}\left(\underset{\omega 1_k}{\longrightarrow}, \underset{\omega 2_k}{\longrightarrow}\right)\right]_{w_k \in d_j} \quad (1)$$

LSTM is used in learning to capture words' past and present informational context. To capture the sequence context by passing forward and backward, we used the bidirectional mode of LSTM (BiLSTM). Then, the output of the BioNER model is a probability of three tags: O, DISO, and ANAT. he positional prefix of the label (B- and I-) is removed, and consideration is given only to the main tag as the correct answer, also referred to as a relaxed match. As the output, a SoftMax activation function is applied to a fully connected layer to calculate the tag probability for each word. We also compare CRF as an alternative in the fully connected layer to incorporate word dependencies as a graphical abstraction.

As LSTM comparison, BERT and XLM-R are transformers-based encoder models to solve NER tasks. Since XLM-R has been trained in more than 100 languages, it gives cross-lingual encoding ability more than BERT. The transformers-based model does not require a separate embedding layer because the tokenizer and encoding process is embedded using sub-word tokenization. For example, an entity "dizzies" is tokenized into ["di", "##zzi", "##es"] that share the same tags ["DISO", "DISO", "DISO"]. As a result, from Fig. 1, our investigations on six BioNER models are (a) Word2Vec-BPEmb + BiLSTM, (b) Flair + BiLSTM, (c) Word2Vec-BPEmb + BiLSTM + CRF, (d) BERT (fine tuning), (e) XLM-R (fine tuning), and (f) BERT + BiLSTM + CRF.

3.3 Semi-supervised learning

Developing high-quality data for a BioNER model requires substantial annotation, i.e., for tens of

$\text{BNER}_{\text{model}}.\text{fitting}(D_{\text{L}})$
while $n_{iter} > 0$ and $|D_{\text{UL}}| > 0$
  $D_{\text{ULS}} \leftarrow \text{sampling}(D_{\text{UL}})$
  $\{\hat{y}_j\} \leftarrow \text{BNER}_{\text{model}}.\text{predicting}(D_{\text{ULS}})$

  for each d_j , if $\text{filtering}(\hat{y}_j, \text{CT})$ then
    $D_{\text{L}} \leftarrow D_{\text{L}} + \langle d_j, \hat{y}_j \rangle$ and $D_{\text{UL}} \leftarrow D_{\text{UL}} - d_j$

  $\text{BNER}_{\text{upd}} \leftarrow \text{BNER}_{\text{model}}.\text{fitting}(D_{\text{L}})$
  $\text{evaluating}(\text{BNER}_{\text{model}}, \text{BNER}_{\text{upd}})$ and dec n_{iter}

Figure 2. Self-training pseudocode for our BioNER model

thousands of documents and more than ten annotators [20, 21]. Semi-supervised learning could improve the BioNER model by giving pseudo labels from the previously trained model on the unannotated data pool to add more training data in an unsupervised setting. We define two datasets of annotated data D_{L} as data training in function $\text{BNER}_{\text{model}}.\text{fitting}(D_{\text{L}})$ with sets of documents $\{d_j\}$ and the corresponding tags of named entities $\{y_j\}$, plus the not annotated data $D_{\text{UL}} = \{d_j\}$.

$$D_{\text{L}} = \{d_j, y_j\} = \left\{[w_{jk}]_{w_{jk} \in d_j}, [tag_{jk}]_{tag_{jk} \in \text{O, DISO, ANAT}}\right\}$$

$$D_{\text{UL}} = \{d_j\} = \left\{[w_{jk}]_{w_{jk} \in d_j}\right\}$$

We use stratified sampling to ensure that all categories of OHC texts in the annotated data have their representatives and are being augmented. To ensure the qualified data is added to retrain the model, a filter function is applied to refine pseudo-labelled data with a particular threshold (Fig. 2). In predicting function, each word will be assigned with its highest probability value of predicted entities. Thus, using function $\text{BNER}_{\text{model}}.\text{predicting}(D_{\text{ULS}})$, the value of $\hat{y}_j$ contains information of named entities tags and the probability values. In function $\text{filtering}(\hat{y}_j, \text{CT})$, for each document d_j with its pseudo labels of named entities $\hat{y}_j$, if the probability values of named entities (NEs) aside of "O" are being averaged and greater than a confidence threshold, CT, then the document d_j will be augmented to the annotated data D_{L}. After that, the retraining process does happen. This mechanism ensures that only high-confidence NEs tag are added to the training dataset. Thus, noise during pseudo-labelling could be minimized. All pseudo-labelled data will not be guaranteed to be augmented to the training dataset. If the evaluation score of new model is higher than the previous model, then the BioNER model is updated accordingly in $\text{evaluating}(\text{BNER}_{\text{model}}, \text{BNER}_{\text{upd}})$.

Table 1. Alodokter data grouped by WHO risks, years, and topics

Year	Top Discussion Topics	# Data
2014-2015	tuberculosis, renal infection, nephrolithiasis	501
2016-2017	***tuberculosis, HIV/AIDS, diarrhea***	3,459
2018-2019	tuberculosis, pneumonia, HIV/AIDS	1,800
2020	covid-19, tuberculosis, bronchitis	2,475
2014-2015	sexuality, fungus infection, men's health	1,748
2016-2017	***pregnancy, menstruation, contraception***	15,423
2018-2019	baby, drugs, pregnancy	9,854
2020	pregnancy, skincare, women health	3,792

4. Results and discussions

4.1 Data preparation

We used OHC texts of Indonesian as a sample of low-resource language for our empirical experiments since it is native to us. The proposed procedures apply to any language that uses the Latin alphabet. We used a corpus of online health-consultation texts scraped from Alodokter.com focusing on the answer parts, and we refer to the corpus as AloData[2]. Texts of questions and answers (QA) are asked by users, answered by authorized medical doctors, and have peer experts to improve their trustworthiness. Alodokter as an OHC platform was selected because of its characteristics in providing informational health support to users.

The dataset contains 325,704 consultation texts with 1,004 categories. Table 1 shows the number of posts and their categories. They are selected based on the WHO categorization of high-risk[3] (i.e., HIV, tuberculosis, diarrhea as coloured rows) or low-risk categories. Bi-yearly aggregate documents indicate that 2016–2017 was the most active usage of this platform, with a total of 155,630 inquiries (± 48% in our data). Discussion topics of tuberculosis (TBC), HIV/AIDS, and diarrhea prompt the most-asked questions in the high-risk category.

[2] https://data.mendeley.com/datasets/p8d5bynh3m

[3] World Health Organization (WHO), "The top 10 causes of death," https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2020.

Table 2. Evaluation results for BioNER comparison based on our procedures

No	Model	Prec.	F1-macro
1	Word2Vec(or W2V) - BPEmb + BiLSTM (SoftMax)	0.7060	0.6989
2	Word2Vec-BPEmb + BiLSTM (CRF)	0.7335	0.7228
3	Flair + BiLSTM (SoftMax)	0.7249	0.6959
4	BERT + BiLSTM (CRF)	0.7616	0.7504
5	BERT + fine tuning	0.8075	0.8022
6	XLM-R + fine tuning	0.8214	0.8152
7	CollaboNET - character-level embedding + BiLSTM (CRF)	0.7281	0.7189
8	CrossNER - BioBERT + BiLSTM (CRF)	0.8173	0.8123

In 2020, there were numerous discussions related to COVID-19. We explored the annotated 2,600 data based on four top high-risk topics (650 tuberculosis, 650 HIV/AIDS, 650 diarrhea, and 650 nephrolithiasis).

The annotated data is split into 75% training (1950), 10% evaluating (260), and 15% testing (390). Four medical experts were assigned to annotate the data using disorder (DISO) and anatomy (ANAT) tags from UMLS guidelines.

Thus, the annotated data [4] has 734,793 tags (including "O" tag), entities with 42,947 DISO and 11,335 ANAT; also, the average words per document is 253.5. The 20,000 unannotated data were selected using a yearly stratified sample on four top high-risk topics to match data distribution for semi-supervised training.

Based on Fig. 1, some models are constructed to assess the best model for BioNER in our case using texts of low-resource language (Table 2) and models from other works on our data, which are CrossNER [8] and CollaboNET [9]. Scenarios 7 and 8 used other models as comparisons, which were executed in our data.

Hyper-parameters for deep learning architectures are as follows.

- *BiLSTM-based models*: (Scenario 1, 2, 3, 4, 7, 8 in Table 2).

Optimal values were found using a stochastic-based Hyperopt Python library [22] with 50 epochs, 128 hidden nodes, learning rate 0.2, dropout rate: 6.797e-05, and batch size 32.

- *Transformers-based models*: (Scenario 4, 5, 6 in Table 2)

Optimal values were following previous work [23] with 20 epochs, the addition of learning rate of 1e-5, AdamW optimizer, and a batch size of 32.

Table 2 shows BioNER identification of annotations from our data validation (260 data) after creating the model with the data training (1950 data). There are around ±70K tag annotations which ±4000 is the DISO tag and ±1000 is the ANAT tag. The testing for Table 2 checked all ±70K tag annotations of those 260 data to get precision and F1 score.

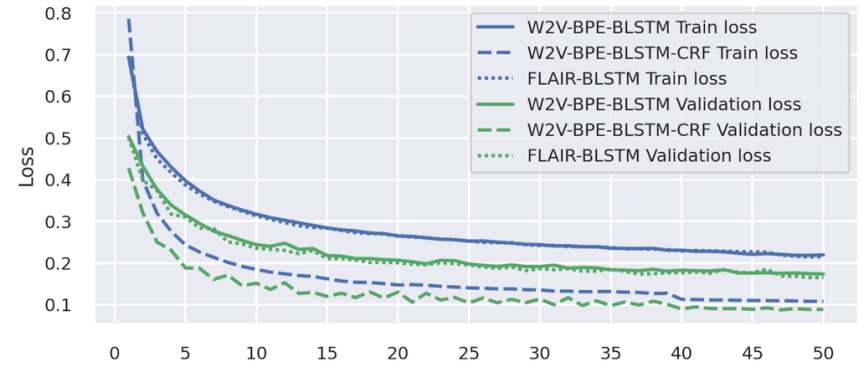
Aside from precision, a macro F1 is used to select the best score since each class is assumed equally important. For BiLSTM-based models, Table 2 shows that additional sub-word embedding of BPE and CRF usage does increase the performance for the NER task (Scenario 2). Instead of the sub-word level, the character level embedding also has an equivalent result (CollaboNET in Scenario 7). Those types of embedding are used to overcome the OOV problem.

However, the transformer-based models have shown higher values for OHC texts that tend to have longer sentences (Scenario 4, 5, 6). We used an existing NLP framework of SparkNLP.

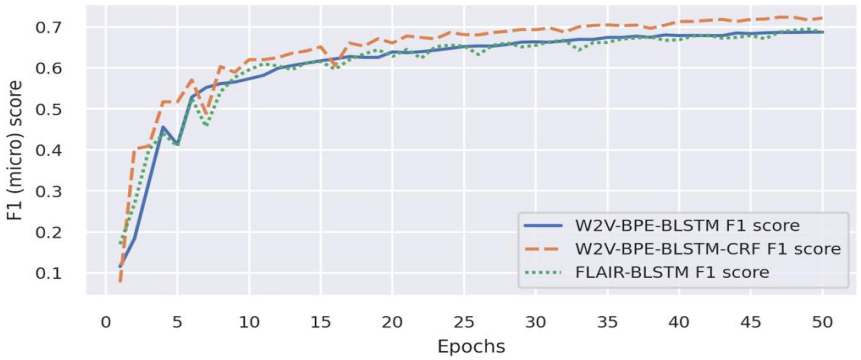
Fig. 3 shows results such that we can better understand how BiLSTM and Transformers models behave during training. The models with a CRF layer gain steeper learning for the training and validation phase, leading to higher training accuracy to show its advantage over SoftMax. All transformer-based models achieve a higher initial loss score, followed by a rapid decline for BERT and XLM-R. Their loss and F1 scores validated that pre-trained stages in transformer-based models help to provide better initial weights since the curves generalize faster (Scenario 4, 5, 6). There is no sign of an irregular loss curve occurring, which means the dataset has equal distribution between the training and test datasets. Interestingly, this occurrence is not present in the hybrid BERT + BiLSTM + CRF model. This behaviour may derive from the integration of BiLSTM layers that make complicated forward and backward relations between the final states of the transformer model.

Our experiments only used datasets in Indonesian texts. However, as discussed, we demonstrated that our precision values on CrossNER and CollaboNET (Table 2) are comparable to other BioNER datasets tested with CrossNER and CollaboNET. Those works of [8, 9] had ranging precision values and had

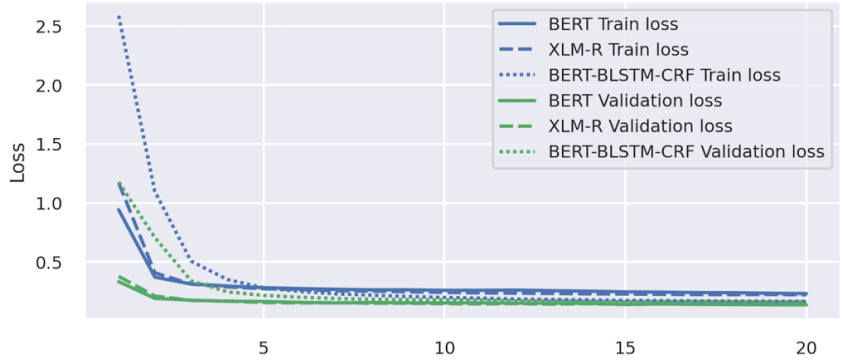
[4] with Labelstudio (github.com/heartexlabs/label-studio)



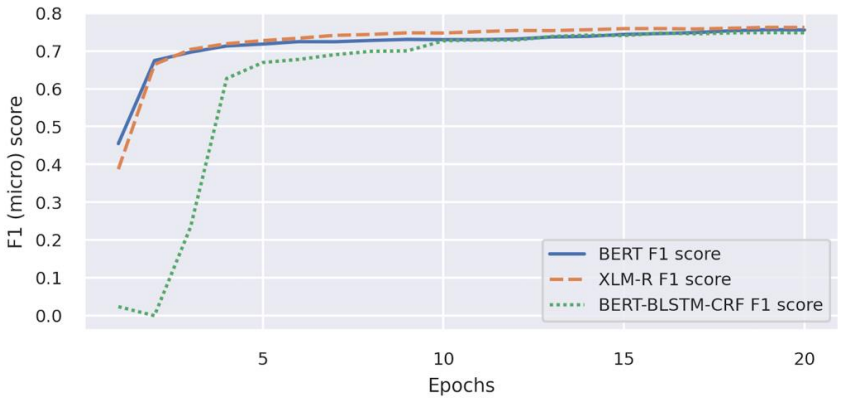
(a)



(b)



(c)



(d)

Figure. 3 Loss (lower values indicate better models) and evaluation (higher values indicate better models) scores of BiLSTM-based and transformer-based models: (a) training loss for BiLSTM model (Epoch vs Loss), (b) F1 scores for BiLSTM model (Epoch vs F1 scores), (c) training loss for transformer model (Epoch vs Loss), and (d) F1 scores for transformer model (Epoch vs F1 scores)

Table 3. Self-training parameters search with values of confidence threshold (CT) to filter eligible data for additional training data

Scen	Model	F1 macro values with different CTs						Δ F1	avg. exec time (mins.)
		Base	CT *0.70*	CT *0.75*	CT *0.80*	CT *0.85*	CT *0.90*		
2	Word2Vec-BPE + BiLSTM-CRF	0.723	0.706	0.712	0.724	**0.732**	0.732	+0.0089	65
6	XLM-R + fine tuning	0.815	0.808	0.821	0.836	**0.838**	0.838	+0.0230	180
7	CollaboNET	0.719	0.699	0.700	0.716	**0.717**	0.710	-0.0015	83
8	CrossNER	0.812	0.800	0.802	0.816	0.822	**0.836**	+0.0238	126

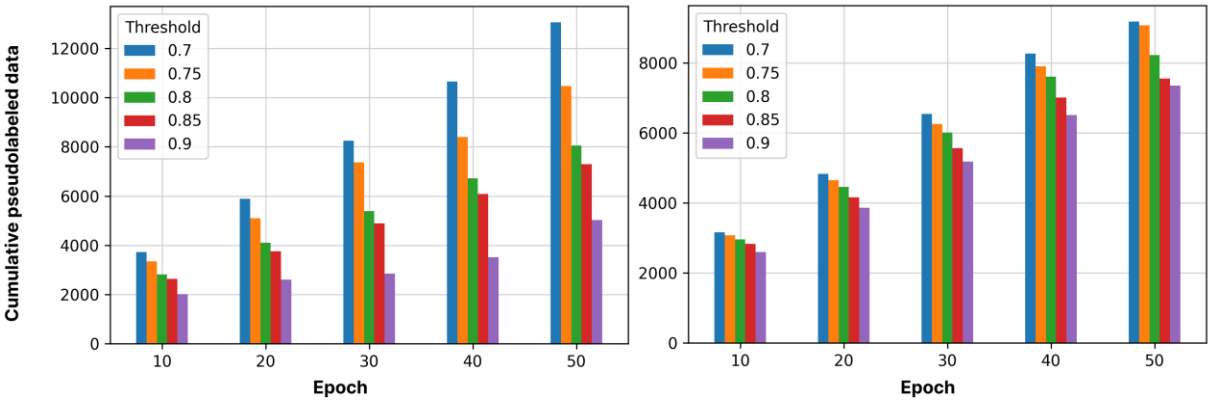


Figure. 4 Cumulative pseudo-labelled data added to training data for each CT value for every tenth epoch. The left side is average data added in BiLSTM model, while the right side is in transformer model

been tested with other BioNER datasets with lots of sentences and annotations compared to our data. However, their lowest score was similar to our result in Table 2 (Scenario 7 and 8) despite record limitations in our datasets. It demonstrated that our approach (Fig. 1) to build low-resource datasets (2600 data) could be followed to gain similar results with far more complete data sources (around 13,000 data in [8] and around 22,000 data in [9]).

4.2 Semi-supervised learning

Our empirical experiments on variations of BioNER using OHC texts of low-resource language of Indonesian have confirmed that the recommended architectures are Word2Vec-BPE + BiLSTM-CRF for BiLSTM based and XLM-R + fine tuning for transformer-based. Subsequent observations are about a semi-supervised learning approach for self-training tasks to improve the BioNER. Table 3 shows the changes in F1 scores with confidence thresholds (CT) to filter out low-quality augmented data.

In BioNER of LSTM-based models (i.e., CollaboNET and Word2Vec-BPE + BiLSTM + CRF), the semi-supervised learning could not increase the performance compared to transformer-based. Transformer-based models need longer training times to test each CT scenario (an average of 121 minutes compared to BiLSTM that requires 74 minutes). Despite the longer training times, still there is an improvement in BioNER models for the transformer-based which measured as $\frac{\Delta F1}{execution\ time}$ (as shown in Table 3).

In BiLSTM-based models of Fig. 4 (left), a large amount of pseudo labelled data was added when CT=0.7, but no increment in CT value, from 0.8 to 0.9. For example, ±3,000 augmented data were generated between the 10th and 50th epochs in the CT=0.9. At the same time, there were ±9,000 additional data produced for a CT value of 0.7. These differences showed that the confidence value of each predicted tag in BiLSTM-based models was 0.7 to 0.8. For Transformer-models in Fig. 4 (right), added data were steady between epochs, meaning there were slight differences data between CT value of 0.7 and 0.9. Total data added with transformer-models in CT 0.9 are also higher than BiLSTM models. It indicates that self-training on OHC-texts with low-resource language performs better with transformer-models

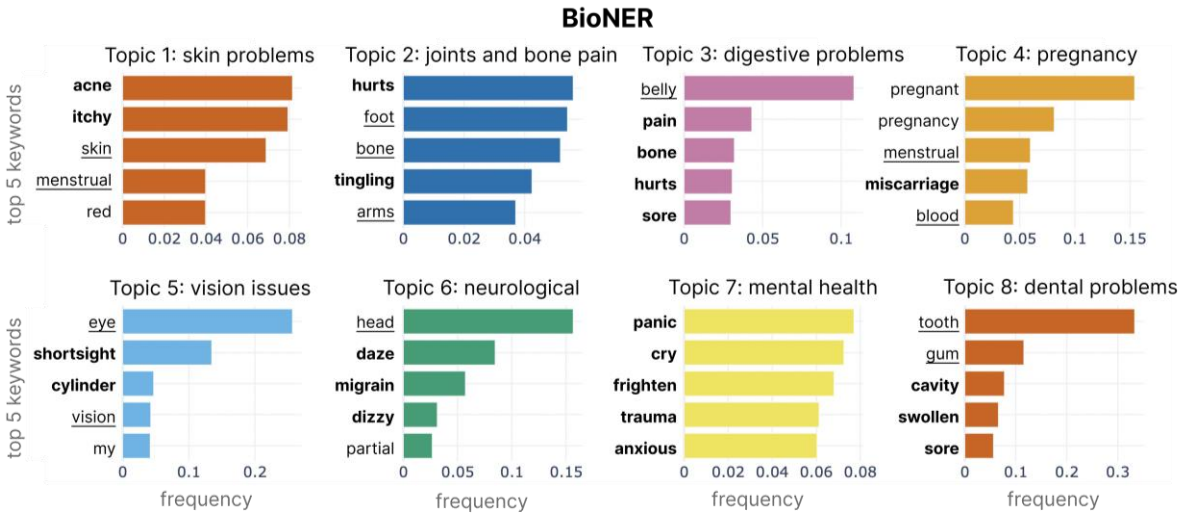


Figure. 5 Extracted entities of DISO and ANAT from OHC texts using the final BioNER model (XLM-R + fine tuning with self-training and CT 0.85). Notes that the original texts are in Indonesian languages as an example of low-resource language

than with BiLSTM models.

4.3 Discussion

These experiments have demonstrated the ability of the proposed framework to extract relevant entities from OHC data. The XLM-R + fine tuning BioNER model achieves the highest score compared to others (± 0.84 in Table 3 scenario 6). Moreover, integrating semi-supervised learning into the framework's pipeline made it possible to achieve a performance gain compared to the original BioNER model (± 0.82 in Table 3 scenario 6).

It was also found that, although trained in limited symptom categories (HIV-AIDS, tuberculosis, nephrolithiasis, and diarrhea), our model could generalize to extracting entities in other categories, such as cancer, vision, and pregnancy problems (as shown in Fig. 5). We used OHC texts from January to December 2021 to demonstrate the final self-trained model and extracted 122,120 DISO and ANAT entities. Those entities become topic keywords as filters to make users easily search the OHC older posts. To understand the topics, we empirically observed those entities after being clustered with several numbers of clusters ranging from 10 to 100. The results showed that ten clusters of identified entities from OHC texts with a Silhouette coefficient of 0.76 gave better word coherence.

Some of those clusters with their top keywords, translated in English, based on weight values of term frequency are listed in Fig. 5. The topic labels are manually defined according to the terms of identified entities as topic keywords. There are some context relations between the topic labels and the topic categories in Table 1. Despite being trained only in specific symptom categories (diarrhea, TBC, HIV-AIDS, and nephrolithiasis or kidney stone), the BioNER model in this study could recognize some words as DISO or ANAT of related terms of never-been-trained categories. To validate this finding, we randomly select 15 questions in OHC for six topics outside of the training categories and evaluate the model performance by doing NER for each question topic. Our findings indicate that entities within some categories have yet to be identified. Specifically, our model recognizes entities of cancer, pregnancy, and vision problems but fails to identify terms of mental health and reproductive issues. It also showed that topics in Fig. 5 comprised the most significant discussions, fluctuating between 40 to 130 monthly posts in the OHC. As a sample use case in the introductory of this work, local government can gain insight through topic keywords of OHC platforms, such that they could set up health promotion activities in public health centres for older adults or pregnant women's self-management because those topics become public concerns.

The proposed framework can support event-based surveillance on public health monitoring [24] by proactively gathering public resources that report symptoms as prominent health signals. Our study could have practical implications for the applicability of the event-based surveillance approach. It will benefit government stakeholders and policymakers as an alternative method of timely intervention in public health events.

5. Conclusion

BioNER models on the high-resource language

are often trained with academic texts of English biomedical literature. Our empirical experiments demonstrated BiLSTM-based and transformer-based use in the named entity recognition (NER) tasks with training on non-formal sentences like posts in online health consultation platforms as an alternative caused by data limitation of standardized and annotated medical texts for low-resource language. We have used an NLP framework in our model for an OHC platform to show the potential of deploying our process for texts written in low-resource languages using self-training with fewer annotated data without requiring task-specific solutions. The experiments demonstrated that the transformer-based had outperformed the BiLSTM-based. Although the semi-supervised predictably introduces noise into a dataset, our filtering function has improved the BioNER model's performance to identify named entities of words that could be topic keywords. Using OHC data for identifying trends through topic keywords could give contextual knowledge of emerging symptom identification as an early warning system.

Conflicts of interest

The authors declare no conflict of interest.

Author contributions

Conceptualization, Diana Purwitasari and Abid Famasya Abdillah; *Methodology*, Diana Purwitasari and Abid Famasya Abdillah; *Software*, Abid Famasya Abdillah; *Validation*, Safitri Juanita and Diana Purwitasari; *Data curation*, Abid Famasya Abdillah, Safitri Juanita and Diana Purwitasari; *Writing—original draft preparation*, Diana Purwitasari and Abid Famasya Abdillah; *Writing—review and editing*, Diana Purwitasari, I Ketut Eddy Purnama, and Mauridhi Hery Purnomo; *Visualization*, I Ketut Eddy Purnama and Safitri Juanita; *Supervision*, Diana Purwitasari and Mauridhi Hery Purnomo.

Acknowledgement

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2023.

References

[1] O. Șerban, N. Thapen, B. Maginnis, C. Hankin, and V. Foot, "Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification", *Information Processing & Management*, Vol. 56, No. 3, pp. 1166-1184, 2019.

[2] A. Husnayain, A. Fuad, and E. C. Y. Su, "Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan", *Intl. Journal of Infectious Diseases*, Vol. 95, pp. 221-223, 2020.

[3] K. Lee, K. Hoti, J. D. Hughes, and L. Emmerton, "Dr Google is here to stay but health care professionals are still valued: An analysis of health care consumers' internet navigation support preferences", *Journal of Medical Internet Research*, Vol. 19, No. 6, p. e210, 2017.

[4] S. Chen, X. Guo, T. Wu, and X. Ju, "Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis", *Information Processing and Management*, Vol. 57, p. 102253, 2020.

[5] X. Jing, "The Unified Medical Language System at 30 years and how it is used and published: Systematic review and content analysis", *JMIR Medical Informatics*, Vol. 9, No. 8, p. e20675, 2021.

[6] P. Sen and A. Saffari, "What do models learn from question answering datasets?", In: *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2429-2438, 2020.

[7] Z. Hu and X. Ma, "A novel neural network model fusion approach for improving medical named entity recognition in online health expert question-answering services", *Expert Systems With Applications*, Vol. 223, p. 119880, 2023.

[8] S. Fan, H. Yu, X. Cai, Y. Geng, G. Li, W. Xu, X. Wang, and Y. Yang, "Multi-attention deep neural network fusing character and word embedding for clinical and biomedical concept extraction", *Information Sciences*, Vol. 608, pp. 778-793, 2022.

[9] W. Yoon, C. H. So, J. Lee, and J. Kang, "CollaboNet: collaboration of deep neural networks for biomedical named entity recognition", *BMC Bioinformatics*, Vol. 20, p. 249, 2019.

[10] L. Weber, M. Sänger, J. Münchmeyer, M. Habibi, U. Leser, and A. Akbik, "HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition", *Bioinformatics*, Vol. 37, No. 17, pp. 2792-2794, 2021.

[11] R. Wongso, Meiliana, and D. Suhartono, "A literature review of question answering system using named entity recognition", In: *Proc. of 3rd*

Intl. Conf. on Information Tech., Computer, and Electrical Engineering (ICITACEE), pp. 274-277, 2016.

[12] Y. M. Kim and T. H. Lee, "Korean clinical entity recognition from diagnosis text using BERT", *BMC Medical Informatics and Decision Making*, Vol. 20(Suppl 7), p. 242, 2020.

[13] H. Yang and H. Gao, "Toward sustainable virtualized healthcare: Extracting medical entities from Chinese online health consultations using deep neural networks", *Sustainability*, Vol. 10, p. 3292, 2018.

[14] Z. Wang and H. Guan, "Research on named entity recognition of doctor-patient question answering community based on BiLSTM-CRF model", In: *Proc. of the 2020 IEEE Intl. Conf. on Bioinformatics and Biomedicine (BIBM)*, pp. 1641-1644, 2020.

[15] I. Moreno, E. Boldrini, P. Moreda, and M. T. R. Ferri, "DrugSemantics: A corpus for named entity recognition in Spanish summaries of product characteristics", *Journal of Biomedical Informatics*, Vol. 72, pp. 8-22, 2017.

[16] S. Gao, O. Kotevska, A. Sorokine, and J. B. Christian, "A pre-training and self-training approach for biomedical named entity recognition", *PLoS ONE*, Vol. 16, No. 2, p. e0246310, 2021.

[17] N. Perera, M. Dehmer, and F. E. Streib, "Named entity recognition and relation detection for biomedical information extraction", *Frontiers in Cell and Developmental Biology*, Vol. 8, p. 673, 2020.

[18] B. Heinzerling and M. Strube, "BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages", In: *Proc. of the Eleventh Intl. Conf. on Language Resources and Evaluation (LREC 2018)*, pp. 2989-2993, 2018.

[19] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, and X. Bai, "Named Entity Recognition using BERT BiLSTM CRF for Chinese Electronic Health Records", In: *Proc. of the 12th Intl. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1-5, 2019.

[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale", In: *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8440-8451, 2020.

[20] R. I. Doğan, R. Leaman, and Z. Lu, "NCBI disease corpus: A resource for disease name recognition and concept normalization", *Journal of Biomedical Informatics*, Vol. 47, pp. 1-10, 2014.

[21] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C. H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "BioCreative V CDR task corpus: A resource for chemical disease relation extraction", *Database*, Vol. 2016, p. baw068, 2016.

[22] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: a Python library for model selection and hyperparameter optimization", *Computational Science & Discovery*, Vol. 8, No. 1, p. 014008, 2015.

[23] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In: *Proc. of the 2019 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171-4186, 2019.

[24] S. A. Balajee, S. J. Salyer, B. G. Cramer, M. Sadek, and A. W. Mounts, "The practice of event-based surveillance: concept and methods", *Global Security: Health, Science and Policy*, Vol. 6, pp. 1-9, 2021.