Long Short-Term Memory for Non-Factoid Answer Selection in Indonesian Question Answering System for Health Information

Retno Kusumaningrum*, Alfi F. Hanifah, Khadijah Khadijah, Sukmawati N. Endah, Priyo S. Sasongko
Department of Informatics, Universitas Diponegoro, Semarang, Indonesia

Abstract—**Providing reliable health information to a community can help raise awareness of the dangers of diseases, their causes, methods of prevention, and treatment. Indonesians are facing various health problems partly due to the lack of health information; hence, the community needs media that can effectively provide reliable health information, namely a question answering (QA) system. The frequently asked questions are non-factoid questions. The development of answer selection based on the classical approach requires distinctive engineering features, linguistic tools, or external resources. It can be solved using deep learning approach such as Convolutional Neural Networks (CNN). However, this model cannot capture the sequence of words in both questions and answers. Therefore, this study aims to implement a long short-term memory (LSTM) model to effectively exploit long-range sequential context information for an answer selection task. In addition, this study analyses various hyper-parameters of Word2Vec and LSTM, such as the dimension, context window, dropout, hidden unit, learning rate, and margin; the corresponding values that yield the best mean reciprocal rank (MRR) and mean average precision (MAP) are found to be 300, 15, 0.25, 100, 0.01, and 0.1, respectively. The best model yields MAP and MRR values of 82.05% and 91.58%, respectively. These results experienced an increase in MAP and MRR of 18.68% and 46.11%, respectively, compared to CNN as the baseline model.**

Keywords—Answer selection; health information; long short-term memory; LSTM; question answering

I. Introduction

Providing reliable health information to a community can help raise awareness of the dangers of diseases, their causes, methods of prevention, and treatment. Indonesians are facing various health problems partly due to the lack of health information, including the dangers of smoking, nutritional problems (stunting and obesity), and serious diseases such as heart disease, cancer, and diabetes. Therefore, the community requires media that can provide health information appropriately, namely a question answering (QA) system.

The QA system is a natural language processing (NLP) application that provides specific answers to the questions/queries posed by the user. The QA system is different from a search engine in that the latter will return a set of documents that may contain answers, and users are required to read the documents and search for the exact answers or infer from the set of documents presented. Therefore, the process of finding answers in a QA system is more complex than the process of finding documents presented by a search engine.

Various QA systems have been developed for both the non-Indonesian QA system and the Indonesian QA system. The following QA systems have been developed for non-Indonesian documents: the English QA system [1]–[3], Chinese QA system [4], Spanish QA system [5]–[7], and French QA system [8], [9]. The Indonesian QA system includes QA statistical and linguistic knowledge systems [10], syntactic-semantic processing QA systems [11], [12], QA systems based on machine learning cross-language QA systems [13], pattern matching QA-based systems [14], and pipeline-based cross-language QA systems [15]. In addition, the Indonesian language QA system has been developed for closed-domain QA [16]–[18].

QA systems are differentiated on the basis of the type of questions handled, which are divided into five categories: factoid, non-factoid, yes-no, list, and opinion [19]. Factoid questions have answers in the form of date, quantity, location, person, organisation, and name (in the form of nouns) in addition to the location, person, and organisation categories [13]. Non-factoid questions are those whose answers are generally used to understand something. Non-factoid questions have six categories: question definitions, reasons, methods, degrees, changes, and details [20]. Overall, the Indonesian QA system is still limited to factoid questions, with hardly any non-factoid questions. Related to health information, the types of questions that are commonly encountered are non-factoid questions.

Several studies have been conducted on non-factoid Indonesian QA systems but for non-health data domains. Moreover, these studies generally used a classical approach such as pattern matching and semantic analysis [21], case-based reasoning [16], and similarity score technique [19]. They provide a good performance only when all the patterns of the answer pairs have been defined, making it appropriate only for certain knowledge domains. In addition, the studies were generally implemented for non-factoid questions related to definitions, reasons, and method categories.

Now-a-days, deep learning models have been widely developed for solving several problems using various types of datasets, such as those containing images, signals, and text. Some examples of deep learning implementation using textual data include sentiment analysis [22], [23], machine translation [24], [25], summarisation [26], [27], and QA. A deep learning model can be implemented in a QA system as a model for selecting the exact answer from a set of candidate answers, also known as the answer pool. The deep learning model does not require feature engineering, linguistic tools, or external resources [28]. Feature engineering is the stage wherein representative features, such as term frequency–inverse document frequency (TF-IDF) and bag-of-words, are determined. The linguistic tools are linguistic rules and syntax. The implementation of deep learning in a QA system requires a convolutional neural network (CNN). However, this model cannot capture the sequence of words in both questions and answers. This can be overcome by implementing long short-term memory (LSTM).

Therefore, this study aims to implement an LSTM as a model for selecting non-factoid answers in the Indonesian question answering system (IQAS) for Health Information. As mentioned earlier, the LSTM model has never been implemented for answer selection in the IQAS, neither in a specific data domain nor in the general data domain. Hence, the first step in this approach is to train the word2vec model on a health information corpus obtained from various popular health websites written in the Indonesian language. In addition, this study empirically analyses the effect of Word2Vec hyper-parameters, such the dimensions and the context window size, on the performance of the LSTM model in selecting the right answer to a question. Furthermore, the effect of varying the LSTM hyper-parameters on the performance of the LSTM model as a model for selecting exact answers from an answer candidate pool was studied; thus, we established the best answer selection model.

The contributions of this paper are summarised as follows:

- A pre-trained Word2Vec model for the Indonesian language, specifically on health information.
- An investigation related to the influences of the dimensions and context window size of Word2Vec on the performance of the LSTM model in selecting answers.
- An analysis of the influences of the hyper-parameters on the LSTM model, including the dropout, number of hidden units, learning rate, and margin size, on the performance of the LSTM model for answer selection.
- A pre-trained LSTM model for non-factoid answer selection in the IQAS for health information. Subsequently, it was implemented as a web-based application.

The rest of this paper is organised as follows. Section II describes related work, including a general description of the answer selection task and LSTM in detail. A detailed explanation of the proposed framework is presented in Section III, including descriptions of data collection, training process of Word2Vec, generation of the answer selection model based on the LSTM, and model evaluation. Section IV presents the experimental results. Finally, in Section V, we draw some conclusions from the results.

II. Related Works

A. Answer Selection Task

Answer selection is a subtask of the QA system that performs the process of selecting sentences containing the required information from a set of candidate answers [29]. Answer selection involves not only matching the terms in the question and answer but also finding the same semantic meaning from both the question and answer. Formally, the answer selection problem can be described as follows:

- There is a question q and answer candidate pool $\{a_1, a_2, \ldots, a_s\}$ that contains a set of answer candidates for a particular question.
- The aim of answer selection is to select the best answer candidates from the answer candidate pool.

Therefore, the answer selection task can be formulated as a ranking problem, giving better ranks to answers that are more relevant to the respective question. Some of the ranking function approaches include pointwise, pairwise, and list wise [30]. This study implements a pairwise approach to train the ranking function to give higher scores for correct answers and lower scores for wrong ones.

B. Long Short-Term Memory

The LSTM model is a popular variation of the recurrent neural network (RNN) method. The RNN method is widely used to solve data problems whose order requires attention. The LSTM model overcomes the gradient vanishing problem of the RNN method. In addition, LSTM model is more capable of dealing with the context of long and sequential information. The LSTM model used in this study is the one introduced in [31].

The LSTM model is designed to solve the gradient vanishing problem using a gate mechanism. Its architecture has three gates, namely an input gate $\boldsymbol{it}$, a forget gate $\boldsymbol{ft}$, and an output gate $\boldsymbol{ot}$, and a memory cell $\boldsymbol{Ct}$. The LSTM can add or reduce information into the cell state, which is regulated by the gate. The input gate is responsible for determining new information to be added to the memory cell. The forget gate determines which information will be saved or deleted. Finally, the output gate is responsible for determining the information that will be used as output. Fig. 1 shows the LSTM cells.

The hidden state $\boldsymbol{ht}$ is calculated on the basis of the three LSTM gates. The size of the hidden state is determined by a parameter called the hidden unit. The hidden unit is a parameter in the LSTM that shows the vector dimension of the hidden state $\boldsymbol{ht}$ for each time step. Mathematically, the LSTM model is defined as follows:

$$i_t = \sigma(W_i x(t) + U_i h(t-1) + b_i) \tag{1}$$

$$f_t = \sigma\left(W_f x(t) + U_f h(t-1) + b_f\right) \tag{2}$$

$$o_t = \sigma(W_o x(t) + U_o h(t-1) + b_o) \tag{3}$$

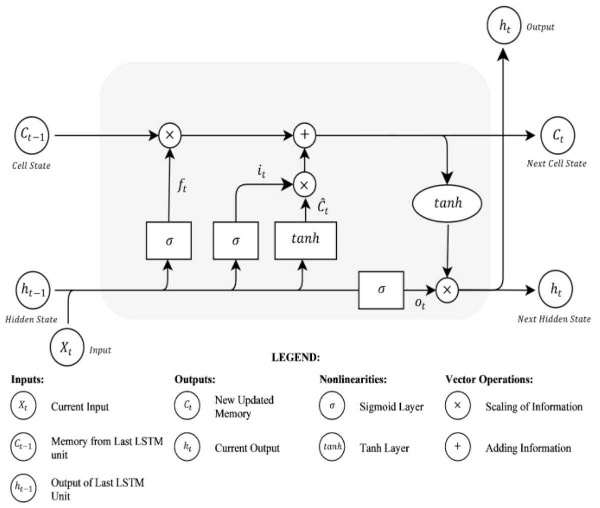


Fig. 1. LSTM cell

$$\tilde{C}_t = tanh(W_c x(t) + U_c h(t-1) + b_c) \quad (4)$$

$$C_t = i_t * \tilde{C}_t + f_t * C_{t-1} \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

The LSTM architecture has three gates (input i, forget f, and output o) and a cell memory vector c. σ is the sigmoid function. W, U, and b are the network parameters.

III. Methodology

This section describes the proposed framework used in this study, comprising four main processes. Fig. 2 shows its general description.

The research framework comprises four main processes: data collection, training process of Word2Vec, generating an answer selection model based on the LSTM, and model evaluation. The detailed explanations for each process are given in the following subsections.

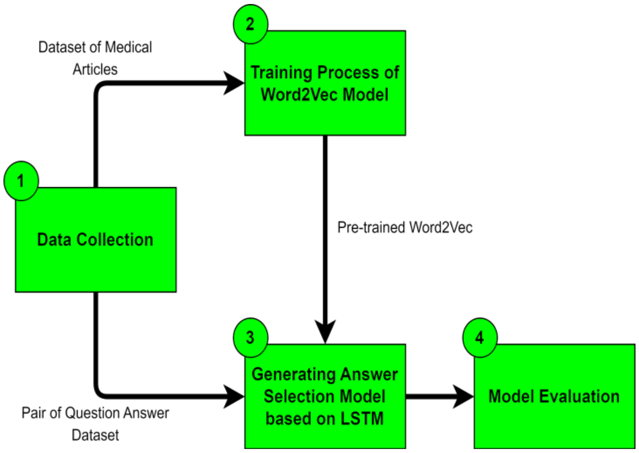

Dataset of Medical Articles
2
Training Process of Word2Vec Model
1
Data Collection
Pre-trained Word2Vec
3
Generating Answer Selection Model based on LSTM
4
Model Evaluation
Pair of Question Answer Dataset

Fig. 2. Framework of this study comprising four main processes: data collection, word2vec training, LSTM-based answer selection modelling, and model evaluation

A. Data Collection

In this process, two types of datasets are formed: a QA dataset (pair of question-and-answer datasets) and a health article dataset. The QA dataset was created by collecting question and answer pairs from popular health sites in Indonesia, namely hellosehat.com, alodokter.com, and halodoc.com. Non-factoid questions on topics of diseases and medicines are used as questions. The categories of the questions are definitions, reasons, and methods. In total, 750 pairs of questions and answers are formed, consisting of 355 pairs for definitions, 145 pairs for reasons, and 250 pairs for methods. The article dataset is established using all the articles from the three websites through data scraping.

B. Training Process of Word2Vec Model

The Word2Vec model is a word embedding algorithm proposed in [32] to learn vector representations. Vector representations can efficiently capture the semantic meaning of the words represented. The word vector tends to obey the laws of analogy and describe intuition. Words known as synonyms have the same vector in the cosine equation, whereas antonyms have different vectors. Therefore, the representation of words in the vector space is useful for achieving better performance on NLP problems by grouping similar words.

The dataset used in Word2Vec training is the article dataset. The article dataset contains articles on diseases and medicines found on the three sites previously described. The number of vocabularies formed was 44,700. The Word2Vec model used is skip-gram, and the evaluation method is hierarchical softmax. Fig. 3 illustrates the skip-gram architecture.

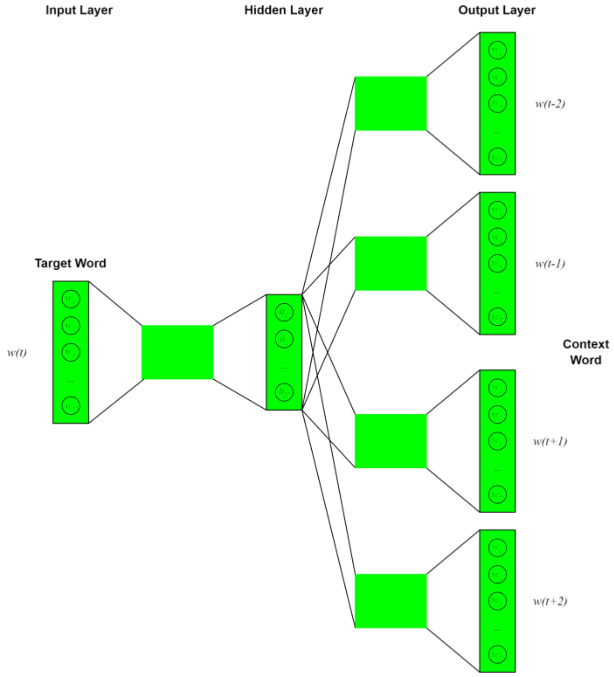


Fig. 3. Illustration of skip-gram architecture of Word2Vec model

C. Generating Answer Selection Model based on LSTM

Modelling for answer selection uses a Siamese architecture. This type of architecture can be used to measure the relevance of candidate answers to a question. Fig. 4 shows the Siamese architecture of the LSTM-based answer selection model. In the embedding layer, the inputted sentences (i.e., the candidate answer and the question) are converted into vector representations generated by Word2Vec training. Thereafter, in the encoding layer, the same encoder is used to create distributed vector representations for the input sentences separately. The encoding layer adopts the QA-LSTM using a bidirectional LSTM (biLSTM) model. During the encoding process, the questions and answers do not have explicit interactions.

Bidirectional LSTM utilises both the previous and future contexts by processing in two directions and generates two independent sequences of LSTM output vectors. The two output vectors are concatenated as follows:

The implementation of max pooling was used to generate representations for the questions and answers based on the word-level biLSTM outputs. The relevance scores of the candidate answers to a question are obtained based on pooled vectors. Subsequently, using the cosine similarity measures the distance between the candidate's answer and the question.

D. Model Evaluation

The evaluation techniques used are the mean reciprocal rank (MRR) and mean average precision (MAP), which are the standard metrics for information retrieval and QA. The MRR can be calculated as follows:

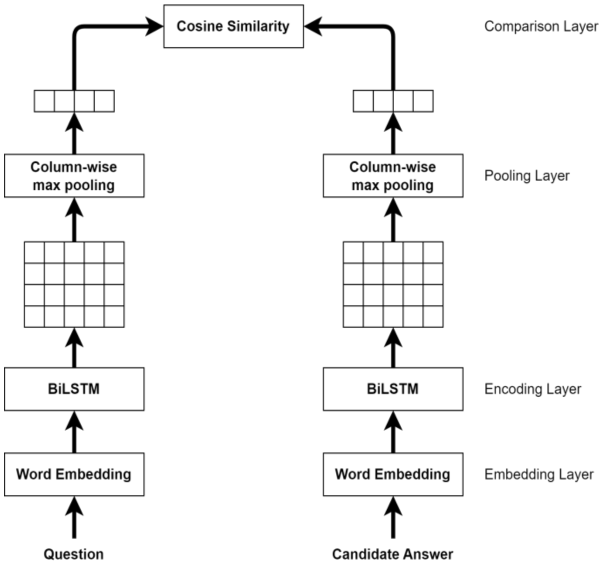


Fig. 4. Siamese architecture of LSTM-based answer selection

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \qquad (7)$$

The MAP can be calculated as follows:

$$MAP = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{|m_j|}Precision(R_{jk}) \qquad (8)$$

IV. Results and Discussion

A. Experimental Setup

The data used in this research are in the form of 750 question–answer pairs. There are 1564 unique answers collected in the answer space. With regard to the distribution ratio of the training and test data, 70% is for training and 30% is for testing. Following the data distribution, we have 525 pairs as training data and 225 pairs as test data. The pool size is 50. It was generated by sending the ground-truth answers to the pool and randomly sampling negative answers from the answer space until the pool size reached 50.

The experiment employs several hyperparameters of the Word2Vec model and LSTM. Each model is trained for 100 epochs. The Word2Vec hyperparameters are dimension (100, 200, and 300) and context window (5, 10, and 15). At the same time, the LSTM hyperparameters are dropout (0.25, 0.5, and 0.75), number of hidden units (50, 75, and 100), learning rate (0.00001, 0.0001, 0.001, and 0.01), and margin (0.05, 0.1, and 0.15).

B. Experimental Scenarios

Several scenarios are established to determine the impacts of the various parameters tested on the performance of the proposed model; scenarios 1, 2, 3, 4, 5, and 6 are for the Word2Vec dimension, context window, dropout, hidden unit, learning rate, and margin, respectively. Fig. 5 shows the overview of these scenarios.

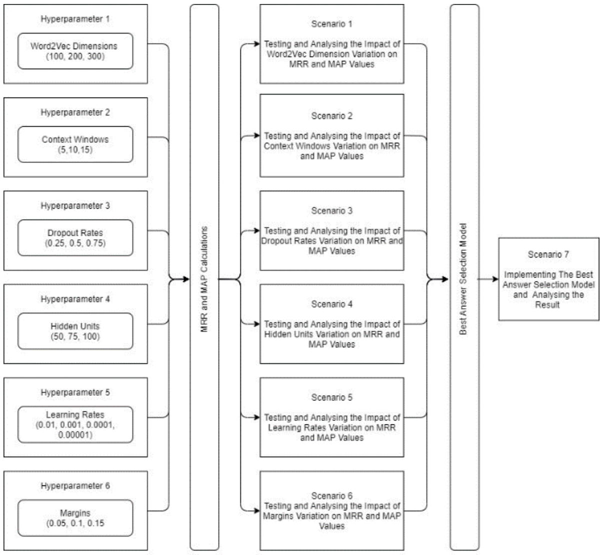

Hyperparameter 1
Word2Vec Dimensions (100, 200, 300)
Hyperparameter 2
Context Windows (5,10,15)
Hyperparameter 3
Dropout Rates (0.25, 0.5, 0.75)
Hyperparameter 4
Hidden Units (50, 75, 100)
Hyperparameter 5
Learning Rates (0.01, 0.001, 0.0001, 0.00001)
Hyperparameter 6
Margins (0.05, 0.1, 0.15
MRR and MAP Calculations
Scenario 1
Testing and Analysing the Impact of Word2Vec Dimension Variation on MRR and MAP Values
Scenario 2
Testing and Analysing the Impact of Context Windows Variation on MRR and MAP Values
Scenario 3
Testing and Analysing the Impact of Dropout Rates Variation on MRR and MAP Values
Scenario 4
Testing and Analysing the Impact of Hidden Units Variation on MRR and MAP Values
Scenario 5
Testing and Analysing the Impact of Learning Rates Variation on MRR and MAP Values
Scenario 6
Testing and Analysing the Impact of Margins Variation on MRR and MAP Values
Best Answer Selection Model
Scenario 7
Implementing The Best Answer Selection Model and Analysing the Result

Fig. 5. Six hyperparameters are tested. Each combination produces a model that calculates the MRR and MAP values. The overall results are analysed through seven scenarios. The best model is obtained from the model that produces the best MRR and MAP values

C. Experimental Results and Analysis

Scenario 1 is aimed at studying the impact of Word2Vec dimensions on the MRR and MAP results. Table I shows that the model yields the best averages of MRR (78.75%) and MAP (63.70%) when the Word2Vec dimension is 300. The MRR and MAP values are directly proportional to the dimensions of Word2Vec; therefore, the higher the dimensions of Word2Vec, the higher the MRR and MAP values. The Word2Vec dimension represents the size of the learned word vector, or it can be referred to as the features of each word. A higher dimension tends to capture more information and better word representations.

Scenario 2 is aimed at studying the impacts of context window on the MRR and MAP results. The best averages of the MRR and MAP values are obtained when the context window is 15, as shown in Table II. From the table, it can be concluded that the averages of the MRR and MAP are directly proportional to the context window, which means that, the larger the context window size, the higher the average MRR and MAP values. The size of the context window defines the range of words to be included as the context of a target word. For instance, a window size of 5 takes five words before and after a target word as its context for training. A larger context window is required to answer non-factoid questions on health information because this type of question requires a longer answer. Moreover, answers related to health information typically have a long explanation.

Scenario 3 is aimed at studying the impacts of dropout rate on the MRR and MAP results. The best MRR and MAP values are 81.25% and 66.58% when the dropout value is set to 0.25. From the average MRR and MAP obtained for all the tested dropout values, it can be concluded that the dropout value is inversely proportional to the average MRR and MAP, which means that, the lower the dropout value, the higher the MRR and MAP. Dropout refers to ignoring units (i.e. neurons) during the training phase of a certain set of neurons. A higher dropout value indicates that more neurons are ignored, and this will cause the model to lose its ability to learn. Moreover, the dropout performed on the LSTM model can make the model to be more limited in keeping the memory. Therefore, lower dropouts are considered better for storing memory in the LSTM model. Table III lists the results of scenario 3.

TABLE I. PERFORMANCE COMPARISON WHEN VARYING THE WORD2VEC DIMENSIONS

Dimension	Average of MAP (%)	Average of MRR (%)
100	56.76	72.91
200	61.77	77.23
300	63.70	78.75

TABLE II. PERFORMANCE COMPARISON WHEN VARYING THE CONTEXT WINDOWS

Context Window	Average of MAP (%)	Average of MRR (%)
5	58.24	74.30
10	61.15	76.63
15	62.84	77.96

TABLE III. PERFORMANCE COMPARISON WHEN VARYING THE DROPOUT RATES

Dropout Rate	Average of MAP (%)	Average of MRR (%)
0.25	66.58	81.25
0.5	62.14	77.50
0.75	53.51	70.14

Scenario 4 is aimed to study the impacts of hidden units on the MRR and MAP results. As mentioned before, this study applies different numbers of hidden units: 50, 75, and 100. From Table IV, it can be concluded that the number of hidden units is directly proportional to the average MRR and MAP. The output dimension determines the number of dimensions for each word in the input sequence. Dimension implies the number of features to be remembered. The best averages of MRR and MAP are obtained under a hidden unit value of 100. This is because using more features provides a better representation than using fewer features.

Scenario 5 is aimed at studying the impacts of the learning rate on the MRR and MAP results. Several learning rates were set: 0.01, 0.001, 0.0001, and 0.00001. Based on Table V, it can be concluded that the learning rate is directly proportional to the averages of MRR and MAP. The best averages of MRR and MAP are obtained under a learning rate of 0.01. As explained in the experimental results section, all the models are trained for 100 epochs. The learning rate is a hyperparameter that helps control the degree of model change. A low learning rate may result in a long training process that could get stuck, making it difficult to converge. These results can be obtained because the epoch used tends to be small; therefore, a high learning rate will decrease the MRR and MAP values.

Scenario 6 is aimed at studying the impact of margin on the MRR and MAP results. As previously explained, there are three different margin values: 0.05, 0.1, and 0.15. The highest average MRR and MAP were obtained under a margin of 0.1, as listed in Table VI. No specific pattern is generated between the margins with the average MRR and MAP. Margin is a variable in the hinge loss function. The hinge loss function is an employed loss function that was minimised in this research. If the ground-truth answer has a score higher than the negative answer by at least a margin, the expression has a zero loss. Condition here implies margins as the optimum distance that can be produced between the ground-truth answer and negative answers. If the margin value is too low, the ground-truth answer and the negative answer will not be separated appropriately. The lower the margin, the smaller the distance between the ground-truth and negative answers. This condition can make relevant answers irretrievable. Meanwhile, if the margin is too high, the distance between the correct answer and the wrong answer will be even greater. This makes irrelevant answers be incorrectly taken as correct answers.

TABLE IV. PERFORMANCE COMPARISON WHEN VARYING THE HIDDEN UNITS

Hidden Units	Average of MAP (%)	Average of MRR (%)
50	58.16	74.15
75	61.12	76.62
100	62.95	78.11

TABLE V. PERFORMANCE COMPARISON WHEN VARYING THE LEARNING RATES

Learning Rates	Average of MAP (%)	Average of MRR (%)
0.00001	50.74	68.41
0.0001	54.14	70.95
0.001	62.56	78.10
0.01	75.53	87.72

TABLE VI. PERFORMANCE COMPARISON WHEN VARYING THE MARGINS

Margins	Average of MAP (%)	Average of MRR (%)
0.05	60.80	76.39
0.1	61.02	76.55
0.15	60.41	75.94

Based on the results of scenarios 1 to 6, the best answer selection model is obtained when using the following hyperparameters: word2vec dimension is 300, context window size is 15, dropout rate value is 0.25, number of hidden units is 100, learning rate is 0.01, and margin value is 0.1. This model yields MAP and MRR values of 82.05% and 91.58%, respectively.

Compared with previous research, this study also run experiments using CNN with an architecture consisting of 4 convolution layers (kernel size in 1, 2, 3, and 5) and one pooling layer. The word2vec dimension used in the test uses the same dimension, namely 300. The best parameter results for the CNN model include margin 0.15, hidden unit 100, dropout 0.25, learning rate 0.01, and context window 15. The MAP and MRR values obtained are 63.37% and 45.47%, respectively. An illustration of the difference between the CNN model and the proposed model can be seen in Fig. 6. It can be seen that the increases in MAP and MRR were 18.68% and 46.11%, respectively.

Subsequently, the best model is implemented for the QA application, which is given the name MediQA. Fig. 7 shows the sample result of the answer selection.

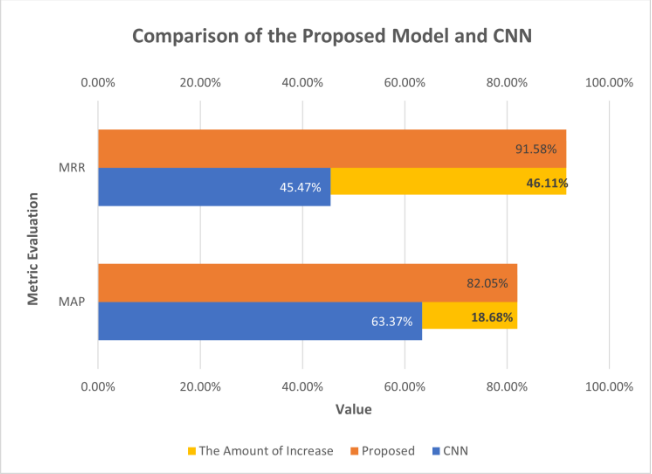
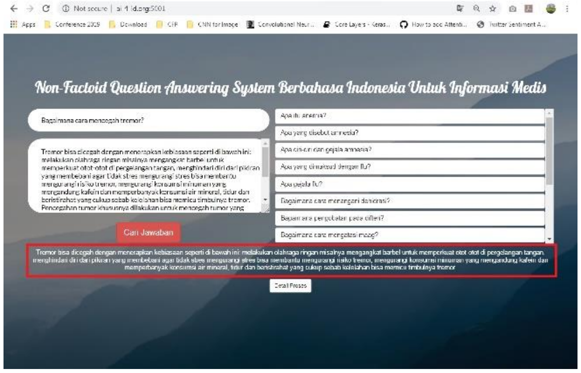


Fig. 6. Comparion of the proposed model and CNN (as baseline model)



Non-Factoid Question Answering System Berbahasa Indonesia Untuk Informasi Medis

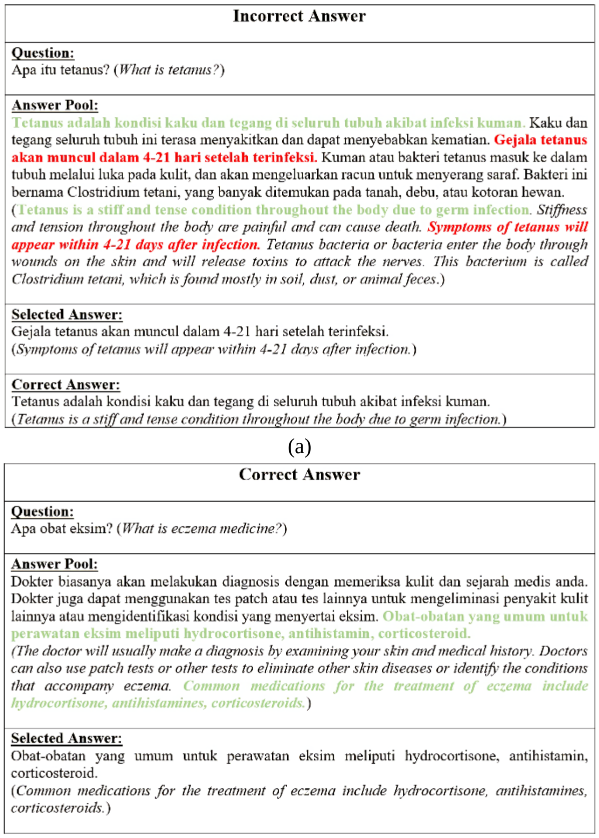
Bagaimana cara mencegah tremor?

Tremor bisa dicegah dengan menerapkan kebiasaan seperti di bawah ini: melakukan olahraga ringan misalnya mengangkat barbel untuk memperkuat otot-otot di pergelangan tangan, menghindari diri dari pikiran yang membebani agar tidak stres mengurangi stres bisa membantu mengurangi risiko tremor, mengurangi konsumsi minuman yang mengandung kafein dan memperbanyak konsumsi air mineral, tidur dan beristirahat yang cukup sebab kelelahan bisa memicu timbulnya tremor. Pencegahan tumor khususnya dilakukan untuk mencegah tumor yang

Apa itu anemia?
Apa yang disebut amnesia?
Apa ciri-ciri dan gejala amnesia?
Apa yang dimaksud dengan flu?
Apa gejala flu?
Bagaimana cara menangani dehidrasi?
Bagaimana pengobatan pada difteri?
Bagaimana cara mengatasi maag?

Tremor bisa dicegah dengan menerapkan kebiasaan seperti di bawah ini: melakukan olahraga ringan misalnya mengangkat barbel untuk memperkuat otot otot di pergelangan tangan, menghindari diri dari pikiran yang membebani agar tidak stres mengurangi stres bisa membantu mengurangi risiko tremor, mengurangi konsumsi minuman yang mengandung kafein dan memperbanyak konsumsi air mineral, tidur dan beristirahat yang cukup sebab kelelahan bisa memicu timbulnya tremor

Fig. 7. Siamese architecture of LSTM-based answer selection



Incorrect Answer

Question:
Apa itu tetanus? (*What is tetanus?*)

Answer Pool:
Tetanus adalah kondisi kaku dan tegang di seluruh tubuh akibat infeksi kuman. Kaku dan tegang seluruh tubuh ini terasa menyakitkan dan dapat menyebabkan kematian. **Gejala tetanus akan muncul dalam 4-21 hari setelah terinfeksi.** Kuman atau bakteri tetanus masuk ke dalam tubuh melalui luka pada kulit, dan akan mengeluarkan racun untuk menyerang saraf. Bakteri ini bernama Clostridium tetani, yang banyak ditemukan pada tanah, debu, atau kotoran hewan. (**Tetanus is a stiff and tense condition throughout the body due to germ infection.** *Stiffness and tension throughout the body are painful and can cause death.* ***Symptoms of tetanus will appear within 4-21 days after infection.*** *Tetanus bacteria or bacteria enter the body through wounds on the skin and will release toxins to attack the nerves. This bacterium is called Clostridium tetani, which is found mostly in soil, dust, or animal feces.*)

Selected Answer:
Gejala tetanus akan muncul dalam 4-21 hari setelah terinfeksi.
(*Symptoms of tetanus will appear within 4-21 days after infection.*)

Correct Answer:
Tetanus adalah kondisi kaku dan tegang di seluruh tubuh akibat infeksi kuman.
(*Tetanus is a stiff and tense condition throughout the body due to germ infection.*)

(a)

Correct Answer

Question:
Apa obat eksim? (*What is eczema medicine?*)

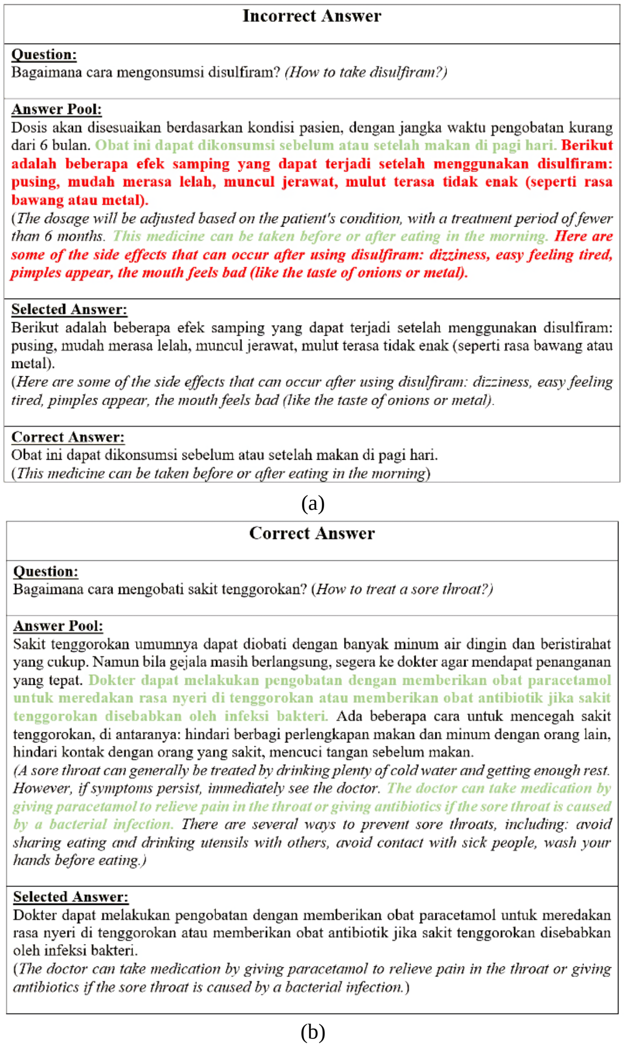
Answer Pool:
Dokter biasanya akan melakukan diagnosis dengan memeriksa kulit dan sejarah medis anda. Dokter juga dapat menggunakan tes patch atau tes lainnya untuk mengeliminasi penyakit kulit lainnya atau mengidentifikasi kondisi yang menyertai eksim. **Obat-obatan yang umum untuk perawatan eksim meliputi hydrocortisone, antihistamin, corticosteroid.**
(*The doctor will usually make a diagnosis by examining your skin and medical history. Doctors can also use patch tests or other tests to eliminate other skin diseases or identify the conditions that accompany eczema. Common medications for the treatment of eczema include hydrocortisone, antihistamines, corticosteroids.*)

Selected Answer:
Obat-obatan yang umum untuk perawatan eksim meliputi hydrocortisone, antihistamin, corticosteroid.
(*Common medications for the treatment of eczema include hydrocortisone, antihistamines, corticosteroids.*)

(b)

Fig. 8. Sample result of answer selection of definition question, (a) Sample of incorrect answer, (b) Sample of correct answer

As mentioned in the previous section, this study evaluates three questions: definitions, reasons, and methods. Fig. 8 shows a sample of the correct and incorrect answer results given by the MediQA application for the definition question type. Fig. 9 shows the same for the method question type. Both figures consist of two parts, the first part shows a result example of choosing the incorrect answer by the system, and the second part shows a result example of choosing the correct answer by the system. In the answer pool section, sentences in green indicate sentences that should have been selected as the correct answer. Meanwhile, sentences written in red are incorrect answer sentences and are output as answers by the system.



Incorrect Answer

Question:
Bagaimana cara mengonsumsi disulfiram? *(How to take disulfiram?)*

Answer Pool:
Dosis akan disesuaikan berdasarkan kondisi pasien, dengan jangka waktu pengobatan kurang dari 6 bulan. **Obat ini dapat dikonsumsi sebelum atau setelah makan di pagi hari.** **Berikut adalah beberapa efek samping yang dapat terjadi setelah menggunakan disulfiram: pusing, mudah merasa lelah, muncul jerawat, mulut terasa tidak enak (seperti rasa bawang atau metal).**
(*The dosage will be adjusted based on the patient's condition, with a treatment period of fewer than 6 months.* ***This medicine can be taken before or after eating in the morning.*** ***Here are some of the side effects that can occur after using disulfiram: dizziness, easy feeling tired, pimples appear, the mouth feels bad (like the taste of onions or metal).***

Selected Answer:
Berikut adalah beberapa efek samping yang dapat terjadi setelah menggunakan disulfiram: pusing, mudah merasa lelah, muncul jerawat, mulut terasa tidak enak (seperti rasa bawang atau metal).
(*Here are some of the side effects that can occur after using disulfiram: dizziness, easy feeling tired, pimples appear, the mouth feels bad (like the taste of onions or metal).*

Correct Answer:
Obat ini dapat dikonsumsi sebelum atau setelah makan di pagi hari.
(*This medicine can be taken before or after eating in the morning*)

(a)

Correct Answer

Question:
Bagaimana cara mengobati sakit tenggorokan? (*How to treat a sore throat?)*

Answer Pool:
Sakit tenggorokan umumnya dapat diobati dengan banyak minum air dingin dan beristirahat yang cukup. Namun bila gejala masih berlangsung, segera ke dokter agar mendapat penanganan yang tepat. **Dokter dapat melakukan pengobatan dengan memberikan obat paracetamol untuk meredakan rasa nyeri di tenggorokan atau memberikan obat antibiotik jika sakit tenggorokan disebabkan oleh infeksi bakteri.** Ada beberapa cara untuk mencegah sakit tenggorokan, di antaranya: hindari berbagi perlengkapan makan dan minum dengan orang lain, hindari kontak dengan orang yang sakit, mencuci tangan sebelum makan.
(A sore throat can generally be treated by drinking plenty of cold water and getting enough rest. However, if symptoms persist, immediately see the doctor. The doctor can take medication by giving paracetamol to relieve pain in the throat or giving antibiotics if the sore throat is caused by a bacterial infection. There are several ways to prevent sore throats, including: avoid sharing eating and drinking utensils with others, avoid contact with sick people, wash your hands before eating.)

Selected Answer:
Dokter dapat melakukan pengobatan dengan memberikan obat paracetamol untuk meredakan rasa nyeri di tenggorokan atau memberikan obat antibiotik jika sakit tenggorokan disebabkan oleh infeksi bakteri.
(*The doctor can take medication by giving paracetamol to relieve pain in the throat or giving antibiotics if the sore throat is caused by a bacterial infection.*)

(b)

Fig. 9. Sample result of answer selection of method question: (a) Sample of incorrect answer, (b) Sample of correct answer

The limitation of this study is that the proposed method focuses on selecting answers on IQAS for a particular domain (health information). At the same time, the need for open-domain QA in Indonesian is still very open. On the other hand, the current state-of-the-art language model reliable for many tasks is Bidirectional Encoder Representations from Transformers (BERT) [33]. The main advantage of BERT is context-sensitive word embedding, where the same word can produce different word embedding when the word has a different context. Word2Vec cannot do this. The Indonesian version of BERT has been developed and is commonly known as IndoBERT [34]. Therefore, it provides an opportunity for further research to apply IndoBERT and LSTM as a model for selecting answers in the Indonesian language open-domain QA.

V. Conclusions

This study analyses various hyperparameters of Word2Vec and LSTM applied to non-factoid answer selection in an IQAS for health information. There are six scenarios to evaluate the effects of the hyperparameters on the MRR and MAP results—first, the larger the dimension of Word2Vec, the better the MRR and MAP values. A dimension of 300 yielded the best MRR and MAP. Second, a context window size of 15 yielded the best MRR and MAP results, indicating that a more extensive context window can yield better MRR and MAP results. Third, a lower dropout value yielded better MRR and MAP values, and the best MRR and MAP were achieved under a dropout value of 0.25. Fourth, the optimum hidden unit value was found to be 100; the higher the number of hidden units, the better the MRR and MAP values. Fifth, a higher learning rate showed significant improvements in the MRR and MAP, given the relatively small number of datasets used in this research. Sixth, a margin of 0.1 produced the best MRR and MAP results. The best model yielded MAP and MRR values of 82.05% and 91.58%, respectively. These results experienced an increase in MAP and MRR of 18.68% and 46.11%, respectively, compared to CNN as the baseline model.

This research is still limited to selecting answers on IQAS for a particular domain (health information), while the need for open-domain QA in Indonesian is still very open. On the other hand, the latest language modelling developments, such as Bidirectional Encoder Representations from Transformers (BERT), have also been developed for Indonesian, commonly known as IndoBERT. Therefore, it provides an opportunity for further research to apply IndoBERT and LSTM as a model for selecting answers in the Indonesian language open-domain QA.

Acknowledgment

This work was supported in part by the Faculty of Science and Mathematics, Universitas Diponegoro under Grant 4866/UN7.5.8/PP/2019.

References

[1] M. Kouylekov and B. Magnini, “Recognizing textual entailment with tree edit distance algorithms,” in *PASCAL Challenges on RTE*, 2006, pp. 17–20.

[2] P. Pakray, S. Pal, S. Bandyopadhyay, and A. Gelbukh, “Automatic answer validation system on english language,” in *ICACTE 2010 - 2010 3rd International Conference on Advanced Computer Theory and Engineering, Proceedings*, 2010, pp. 329–333.

[3] S. K. Ray, S. Singh, and B. P. Joshi, “World wide web based question answering system - A relevance feedback framework for automatic answer validation,” in *2nd International Conference on the Applications of Digital Information and Web Technologies, ICADIWT 2009*, 2009, pp. 169–174.

[4] D. Cai, Y. Dong, D. Lv, G. Zhang, and X. Miao, “A web-based Chinese question answering with answering validation,” in *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE’05*, 2005, pp. 499–502.

[5] Á. Rodrigo, A. Peñas, and F. Verdejo, “The effect of entity recognition in the answer validation,” in *CEUR Workshop Proceedings*, 2006, vol. 1172, pp. 1–5.

[6] A. Téllez-Valero, M. Montes-Y-Gómez, L. Villaseñor-Pineda, and A. Peñas, “Improving question answering by combining multiple systems via answer validation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4919 LNCS, pp. 544–554, 2008.

[7] A. Téllez-Valero, M. Montes-y-Gömez, L. Villaseñor-Pineda, and A. Peñas-Padilla, “Towards multi-stream question answering using answer validation,” *Inform.*, vol. 34, no. 1, pp. 45–54, 2010.

[8] A. L. Ligozat, B. Grau, A. Vilnat, I. Robba, and A. Grappy, “Towards

an automatic validation of answers in question answering," in *Proceedings of International Conference on Tools with Artificial Intelligence, ICTAI*, 2007, pp. 444–447.

[9] A. Grappy, B. Grau, M. H. Falco, A. L. Ligozat, I. Robba, and A. Vilnat, "Selecting answers to questions from Web documents by a robust validation process," in *Proceedings of 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011*, 2011, pp. 55–62.

[10] M. Adriani and S. Adiwibowo, "Finding answers using resources in the internet," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5152 LNCS, pp. 332–335, 2008.

[11] S. D. Larasati and R. Manurung, "Towards a semantic analysis of bahasa Indonesia for question answering," *Proc. 10th Conf. Pacific Assoc. Comput. Linguist.*, pp. 273–280, 2007.

[12] R. Mahendra, S. D. Larasati, and R. Manurung, "Extending an Indonesian semantic analysis-based question answering system with linguistic and world knowledge axioms," in *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, 2008, pp. 262–271.

[13] A. Purwarianti, M. Tsuchiya, and S. Nakagawa, "A machine learning approach for an Indonesian-English cross language question answering system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 11, pp. 1841–1852, 2007.

[14] H. Toba and M. Adriani, "Pattern based Indonesian question answering system," in *Proceedings of the 1st International Conference on Advanced Computer Systems and Information Systems (ICACSIS)*, 2009, pp. 1–6.

[15] M. I. Faruqi and A. Purwarianti, "An Indonesian question analyzer to enhance the performance of Indonesian-English CLQA," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, ICEEI 2011*, 2011, pp. K2-1.

[16] A. Fikri and A. Purwarianti, "Case based Indonesian closed domain question answering system with real world questions," in *Proceeding of the 7th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2012*, 2012, pp. 181–186.

[17] A. A. S. Gunawan, P. R. Mulyono, and W. Budiharto, "Indonesian question answering system for solving arithmetic word problems on intelligent humanoid robot," in *Procedia Computer Science*, 2018, vol. 135, pp. 719–726.

[18] R. H. Gusmita, Y. Durachman, S. Harun, A. F. Firmansyah, H. T. Sukmana, and A. Suhaimi, "A rule-based question answering system on relevant documents of Indonesian Quran translation," in *Proceeding of 2014 International Conference on Cyber and IT Service Management, CITSM 2014*, 2014, pp. 104–107.

[19] N. Yusliani and A. Purwarianti, "Sistem Question Answering Bahasa Indonesia untuk Pertanyaan Non-Factoid," *J. Ilmu Komput. dan Inf.*, vol. 4, no. 1, p. 10, 2012.

[20] M. Murata, S. Tsukawaki, T. Kanamaru, Q. Ma, and H. Isahara, "A system for answering non-factoid Japanese questions by using passage retrieval weighted based on type of answer," pp. 2–7, 2007.

[21] A. A. Zulen and A. Purwarianti, "Study and implementation of monolingual approach on indonesian question answering for factoid and non-factoid question," in *PACLIC 25 - Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 2011, pp. 622–631.

[22] A. S. Zharmagambetov and A. A. Pak, "Sentiment analysis of a document using deep learning approach and decision trees," in *Proceedings of the 2015 12th International Conference on Electronics Computer and Computation, ICECCO 2015*, 2016, pp. 1–4.

[23] M. Y. Day and Y. Da Lin, "Deep learning for sentiment analysis on google play consumer review," in *Proceedings of 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, 2017, pp. 382–388.

[24] J. Zhang and C. Zong, "Deep neural networks in machine translation: An overview," *IEEE Intell. Syst.*, vol. 30, no. 5, pp. 16–25, 2015.

[25] S. P. Singh, A. Kumar, H. Darbari, L. Singh, A. Rastogi, and S. Jain, "Machine translation using deep learning: An overview," in *Proceeding of 2017 International Conference on Computer, Communications and Electronics, COMPTELIX 2017*, 2017, pp. 162–167.

[26] S. P. Singh, A. Kumar, A. Mangal, and S. Singhal, "Bilingual automatic text summarization using unsupervised deep learning," in *Proceeding of International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 1195–1200.

[27] C. Yao, J. Shen, and G. Chen, "Automatic document summarization via deep neural networks," in *Proceedings of 2015 8th International Symposium on Computational Intelligence and Design, ISCID 2015*, 2016, pp. 291–296.

[28] M. Tan, C. dos Santos, B. Xiang, and B. Zhou, "LSTM-based deep learning models for non-factoid answer selection," in *Proceeding of the 4th International Conference on Learning Representation (ICLR 2016)*, 2016, no. 1, pp. 1–11.

[29] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman, "Deep learning for answer sentence selection," 2014. [Online]. Available: https://arxiv.org/abs/1412.1632.

[30] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retr.*, vol. 3, no. 3, pp. 225–331, 2009.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] T. Mikolov, G. Corrado, K. Chen, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013, pp. 1–12.

[33] J. Devlin, M.-W. Chang, K. Lee, and K Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, 2019, pp. 4171–4186.

[34] F. Koto, A. Rahimi, J.H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, 2020, pp. 757–770.