TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 20, No. 6, December 2022, pp. 1248~1255
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v20i6.24248

RoBERTa: language modelling in building Indonesian question-answering systems

Wiwin Suwarningsih[1], Raka Aditya Pratama[2], Fadhil Yusuf Rahadika[2], Mochamad Havid Albar Purnomo[2]
[1]Research Center for Data Science and Information, National Research and Innovation Agency, Bandung, Indonesia
[2]Department of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia

Article Info

Article history:

Received Jul 01, 2021
Revised Sep 09, 2022
Accepted Sep 19, 2022

Keywords:

ALBERT
ELECTRA
Indonesian QAS
Language modelling
RoBERTa

ABSTRACT

This research aimed to evaluate the performance of the A Lite BERT (ALBERT), efficiently learning an encoder that classifies token replacements accurately (ELECTRA) and a robust optimized BERT pretraining approach (RoBERTa) models to support the development of the Indonesian language question and answer system model. The evaluation carried out used Indonesian, Malay and Esperanto. Here, Esperanto was used as a comparison of Indonesian because it is international, which does not belong to any person or country and this then make it neutral. Compared to other foreign languages, the structure and construction of Esperanto is relatively simple. The dataset used was the result of crawling Wikipedia for Indonesian and Open Super-large Crawled ALMAnaCH coRpus (OSCAR) for Esperanto. The size of the token dictionary used in the test used approximately 30,000 sub tokens in both the SentencePiece and byte-level byte pair encoding methods (ByteLevelBPE). The test was carried out with the learning rates of 1e-5 and 5e-5 for both languages in accordance with the reference from the bidirectional encoder representations from transformers (BERT) paper. As shown in the final result of this study, the ALBERT and RoBERTa models in Esperanto showed the results of the loss calculation that were not much different. This showed that the RoBERTa model was better to implement an Indonesian question and answer system.

Corresponding Author:

Wiwin Suwarningsih
Research Center for Data Science and Information, National Research and Innovation Agency
Komplek LIPI Jl. Sangkuriang 21 Bandung 40135, Indonesia
Email: wiwin.suwarningsih@brin.go.id

1. INTRODUCTION

The development of a question answer system (QAS) requires the development of an appropriate language model. This is necessary because an ideal QAS system must have four supporting processes, which include candidate document selection, answer extraction, answer validation, and response generation [1], [2]. The problem that arises is which modelling technique is suitable for selecting candidate documents to increase question answer (QA) performance [3], [4]. A research conducted by Tan *et al.* [5] applied the bidirectional long-short term memory (Bi-LSTM) model for questions and answers that were connected with later pooling to compare similarities. To form a better embedding, a convolutional neural network (CNN) layer was added after Bi-LSTM. In addition, to better separate answer candidates based on questions, an embedding model was introduced for answers based on the context of the question. Similarly, Akbik *et al.* [6] developed standard Word2vec using a Bi-LSTM layer plus character embedding. Here, character embedding was used to handle words that were outside the bag of word.

Handling the validation of answers from the QA architecture is an important spotlight to improve QAS performance because this architecture can be developed as an improvement from the existing architecture [7]. What is different about this QA architecture is that there is an addition to the answer generator. After the answer is obtained, the answer will be used as input again in the answer generator [8], [9]. This answer generator will rearrange the answers obtained with additional input from the feature extraction results from questions using char recurrent neural network (RNN) [10], [11].

The two problems above, namely sorting candidate documents and validating answers have been handled by several methods such as the application of long-short term memory-recurrent neural network (LSTM-RNN) [12], template convolutional recurrent neural network (T-CRNN) [13], CNN-BiLSTM [14], dynamic co-attention networks (DCN) [15]. In [12] the training model used seq2seq for embedding with a learning rate of 0.001 to avoid the gradient disappearing, and to ensure high answer accuracy. In contrast [13] applied the T-CRNN to obtain a right correlation between answers and questions. The model training was carried out using a number of features mapped for each layer, namely 100 features with pre-trained word embedding using 100 dimensions. While in [14] the system tested the collaborative learning based answer selection model (QA-CL) combined with CNN and Bi-LSTM architectures as initial matrix vector initialization. This model applied weight removal (WR) for embedding question sentences and generated the initial vector matrix, and used principal component analysis (PCA) for feature reduction. It not only applied a hybrid model of combining LSTM with CNN, but also applied CNN behind the LSTM layer to study the representation of the question answer sentences. In contrast [15] implemented DCN to correct possible errors in answers due to local maxima in the QA System. The dataset used in this study was the Stanford question answering dataset (SQuAD). Basically, in SQuAD there is an intuitive method that generates an answer index range by predicting the start and end of the index range.

Another way to study the context of a word based on the words around it, not just the words that precede or follow it, allows using a language model also known as transformers' bidirectional encoder representation (BERT) [16]. The breakthrough of BERT is its ability to train language models based on the whole set of words in a sentence or query (two-way training) rather than the traditional way of training on a sequenced word order. Similarly, the research conducted by [17] used a BERT-based fine-tuning approach. The purpose of using BERT is to reduce the constraints when pre-training masked language (MLM) models. To improve the performance in the process of training the model for longer, with larger batches and more data, the variance of BERT emerged such as a robust optimized BERT pretraining approach (RoBERTa) [18] and ALBERT [19].

RoBERTa [18] is the result of a modified BERT pre-training procedure that improves the final project performance. It focuses on large-scale byte and batch increments so that it can be used to train with dynamic masking without losing next sentence prediction (NSP). The use of RoBERTa to explore natural language generation in question-answering [20], focused more on producing short answers to the given questions and did not require any annotated data. The proposed approach showed some very promising results for English and French. Whereas in the study [21] the trees induced from RoBERTa outperformed the trees provided by the parser. The experimental results showed that the language model implicitly incorporated the task-oriented syntactic information.

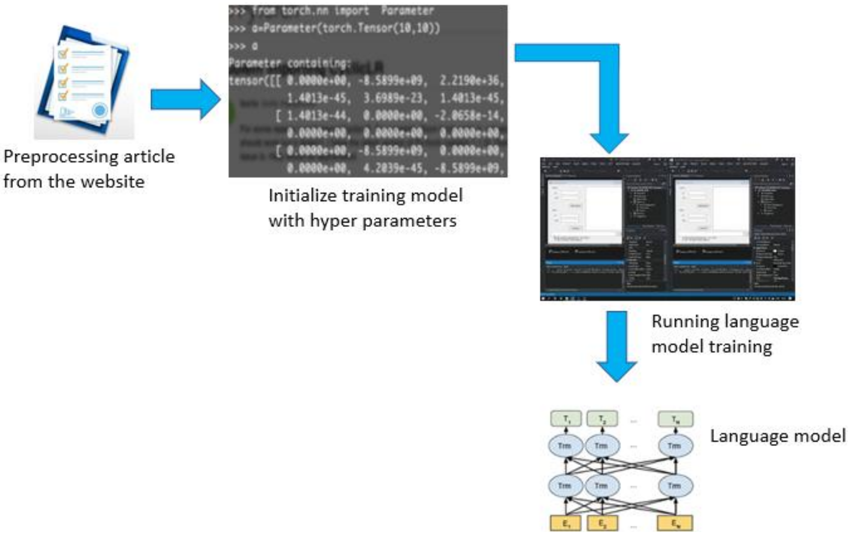
A Lite BERT (ALBERT) for self-supervised learning of language representation model [19] has a much smaller parameter size compared to the BERT model. ALBERT combines two parameter reduction techniques that remove the main bottleneck in scaling the pre-trained model. The method we used in this study was to compare ALBERT and RoBERTa. ALBERT simplifies the BERT architecture by utilizing parameter-sharing and using the SentencePiece tokenization technique in the pre-training stage. RoBERTa made some modifications to the pre-training section of BERT and removed the NSP method in training. RoBERTa uses the byte-level byte pair encoding (ByteLevelBPE) tokenization method adapted from generative pre-trained transformer-2 (GPT-2).

Based on the explanations above, there are still several things that must be addressed, including how to process long sequential data and collect all information from text in network memory. In this study, we proposed RoBERTa – an artificial neural network based on transformers with 6 layers (base model). RoBERTa [22], [23] used the WordPiece tokenization technique in the pre-training stage and the training method used masked layer modelling (MLM) and NSP to support our QAS system. Our contribution are: 1) representing the extraction of language models at the character level for answer selection without any engineering features and linguistic tools; and 2) applying an efficient self-attention model to generate answers according to context by calculating the input and output representations regardless of word order.

The rest of the paper is organized as: section 2 describes research method. The research method contains the stages of work we did in this paper. The result and analysis are shown in section 3, this section describes test results of the method we used. Finally, section 4 presents the conclusion of the paper. In this conclusion section it is explained that we have solved the problem based on the method we use and testing the data we have.

2. RESEARCH METHOD

The research method in this study can be seen in Figure 1. The approach used for training data used ALBERT [19], RoBERTa [18], [21] and efficiently learning an encoder that classifies token replacements accurately (ELECTRA) [24], [25]. The goal was to obtain an ideal model to test with our dataset. The process of evaluating the performance of the model used two languages, i.e. Indonesian and Esperanto. Esperanto was used as a comparison to Indonesian for several reasons, including 1) Esperanto is international, which does not belong to a person or a country. In other words, it is neutral; 2) Esperanto's structure and construction are relatively simple compared to other foreign languages; and 3) Esperanto is fair and everyone has the equal rights.



```
>>> from torch.nn import Parameter
>>> a=Parameter(torch.Tensor(10,10))
>>> a
Parameter containing:
tensor([[ 0.0000e+00, -8.5899e+09,  2.2190e+36,
          1.4013e-45,  3.6989e-23,  1.4013e-45,
        [ 1.4013e-44,  0.0000e+00, -2.0658e-14,
          0.0000e+00,  0.0000e+00,  0.0000e+00,
        [ 0.0000e+00, -8.5899e+09,  0.0000e+00,
          0.0000e+00,  4.2039e-45, -8.5899e+09,
```

Figure 1. Proposed research method

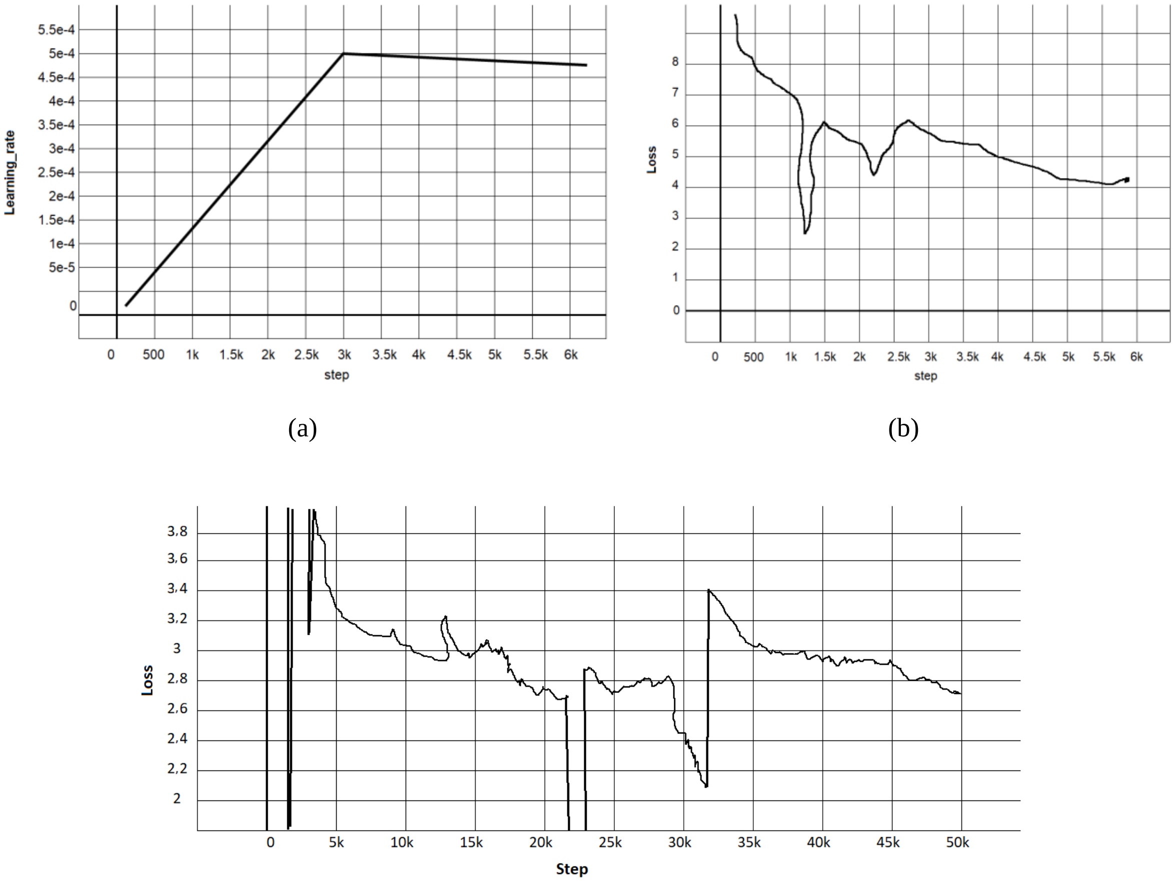
Based on the proposed method in Figure 1, the article about coronavirus disease 2019 (COVID'19) news (we got it from crawling results on Indonesian Wikipedia, Jakarta News, Okezone, Antara, Kumparan, Tribune, and Open Super-large Crawled ALMAnaCH coRpus (OSCAR)) which is the input data for our study in preprocessing and converting the format to be used as input data for our study as a knowledge base system. Then the training model for initialize the parameters which includes the number of layers, optimizer, number of parameters, learning rate, learning rate scheduler, weight_decay, adam_beta1, adam_beta2, warmup step ratio, steps and batch size. This initialization is needed to facilitate the process of running the model language using BERT which consists of RoBERTa, ALBERT, and ELECTRA. The training model in this study using high-performance computing provided by the national research and innovation agency took more than 36 hours to train on a single graphics processing unit (GPU) V100. The final result of the running model language process is the RoBERTa, ALBERT, and ELECTRA language models which will be used for testing documents that are used as answer candidates in the question and answering system that will be built.

The dataset used was the result of crawling Wikipedia for Indonesian and OSCAR for Esperanto. The size of the token dictionary used in the test used approximately 30,000 sub tokens in both the SentencePiece and ByteLevelBPE methods. The test was carried out with the learning rates of 1e-5 and 5e-5 for both languages according to the reference from the BERT paper [17], [22], [26].

3. RESULTS AND ANALYSIS

3.1. Long training RoBERTa

From our previous research experience, RoBERTa showed the better results than ALBERT for Indonesian [27], but the training time tended to be long (restricted resources). Thus, we did long training for RoBERTa with loss results like the graph in the picture above. The graph showed a "trough" caused by a change in the distribution of the corpus. The change in the distribution of the corpus occurred due to the shift from the Wikipedia corpus, which is generally standard language to the OSCAR corpus. The OSCAR corpus is the result of free web scraping by CommonCrawl, which is further classified by language. The corpus has the disadvantage for containing a number of dirty or rude words generally found on gambling websites or free forums.



(a) (b)

(c)

Figure 2. Long training RoBERTa: (a) learning rate graph, (b) loss graph, and (c) performance graph RoBERTa language model

In Figure 2 performance language model, the loss value looks like a drastic drop like the trough phenomenon. Figure 2(a) explain learning rate graph is a hyper-parameter that controls how much we are adjusting the weights of our network with respect the loss gradient. Figure 2(b) loss graph is function results for different models and Figure 2(c) performance graph RoBERTa language model shows that pretraining multilingual language models at scale leads to significant. This was because data had the same structure and it made it simple for the model to guess. An unstable loss value did not indicate that the neural network decreased in capacity. This was because we loaded the corpus data sequentially, not randomly. The reason was because of limited resources so that the corpus was not able to be loaded entirely at one time but broken down into several parts.

3.2. Comparison of ELECTRA with RoBERTa

We are interested in testing the performance of the ELECTRA model as a baseline for the method we use because the SQuAD explorer leaderboard in [15] showed the quite good results. ELECTRA uses a new pre-training task called substitution token detection that trains a bidirectional model while learning from all input positions. ELECTRA trains models to distinguish between “real” and “fake” input data. In ELECTRA, the input is not masked, but it is corrupted by the approach of replacing some input tokens with plausible alternatives obtained from a small generator network. Figure 3 shows the results of ELECTRA and RoBERTa’s learning rate using Malay, which tends to be similar to Indonesian.

Based on the learning rate in Figure 3, the next test was to conduct QA training using the ELECTRA model language of Malaysian language, which tends to be similar to Indonesian. Of 65 Indonesian QA-Pairs datasets manually selected and then formatted, so that they became SQuAD version 1.1, 100 epochs of training were conducted (see Figure 3(a)). The Malaysian ELECTRA model produced a satisfactory performance with an exact match (EM) score of 0.8 and an F1-measure score of 0.84 (see Figure 3(b)).

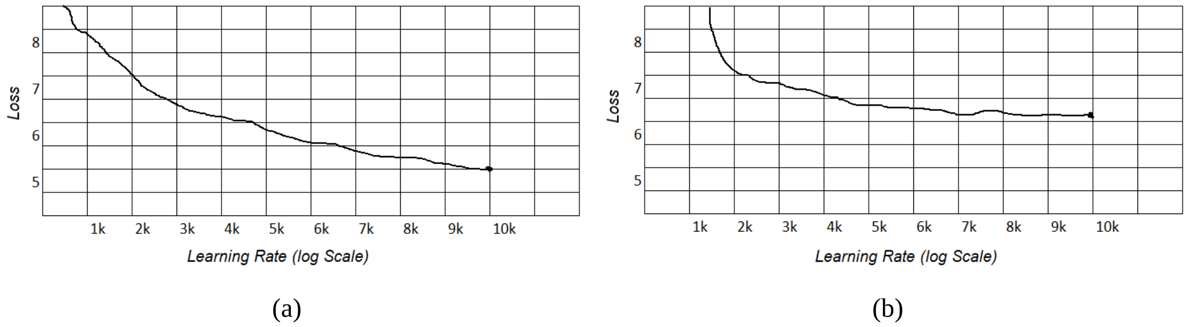


Figure 3. Learning rate RoBERTa vs ELECTRA: (a) RoBERTa with a learning rate of 1e-4 and (b) ELECTRA with a learning rate of 1e-4

3.3. Comparison of ALBERT with RoBERTa

At this stage, the next step was to compare Albert's performance with RoBERTa's. The learning rate we used applied two intervals, namely 1e-5 and 5e-5 [19], [18]. The experimental results can be seen in Figure 4.

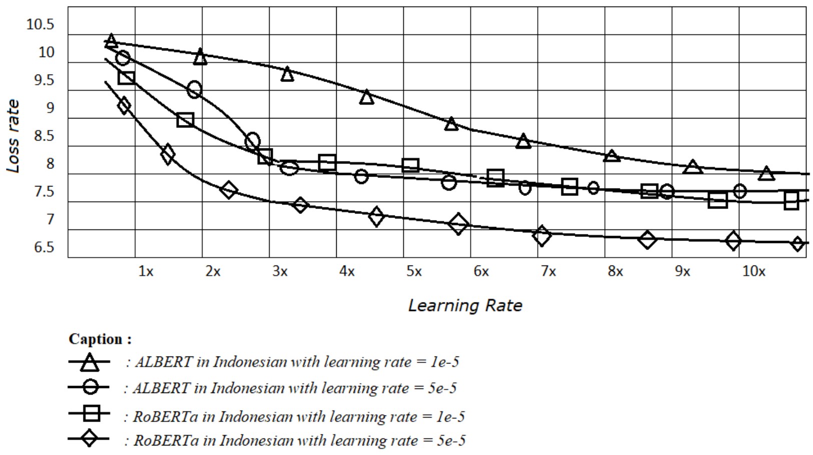


Figure 4. Learning rate 5e-5 ALBERT and RoBERTa

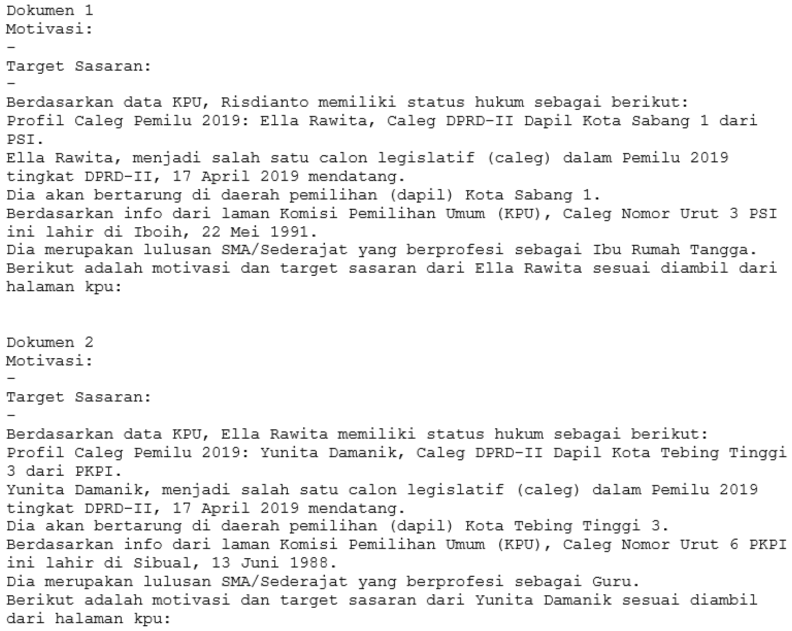
As seen in Figure 4, the ALBERT and RoBERTa models in Esperanto showed the results of the "loss" calculation that were not much different. The RoBERTa model showed a calculation result that tended to be better than ALBERT in Indonesian with a significant "loss" calculation result. The use of learning rate 5e-5 showed the better results than 1e-5.

3.4. Result and discussion

The language model training stage is the training stage where the model is trained to understand the structure of the language used. The training strategy used is to mask some of the tokens (a common percentage or recommended for use is 15% of masked tokens) in the input sentence, and then the model will be trained to predict missing words or masked words. The data used at this stage were the unlabelled data. The result of this training phase was a basic model that only had language skills. The language skills of the next model can be used for knowledge base and retrained (fine-task tuning).

This training stage can be skipped by using a published pre-trained model so that only fine-task tuning was required. When this research was started, no specific model of Indonesian based on transformers has, so far, been published in general, so the researcher decided to train the language model independently. During this research, several types of models were published by other researchers. This model was then used as a comparison against the model that was trained independently in this study.

The experimental results in Figure 2. Shows the occurrence of a "trough" due to the presence of the same structured data so that it was simple for the model to guess. An example of the part of the body that caused the trough can be seen in Figure 5.



```
Dokumen 1
Motivasi:
-
Target Sasaran:
-
Berdasarkan data KPU, Risdianto memiliki status hukum sebagai berikut:
Profil Caleg Pemilu 2019: Ella Rawita, Caleg DPRD-II Dapil Kota Sabang 1 dari
PSI.
Ella Rawita, menjadi salah satu calon legislatif (caleg) dalam Pemilu 2019
tingkat DPRD-II, 17 April 2019 mendatang.
Dia akan bertarung di daerah pemilihan (dapil) Kota Sabang 1.
Berdasarkan info dari laman Komisi Pemilihan Umum (KPU), Caleg Nomor Urut 3 PSI
ini lahir di Iboih, 22 Mei 1991.
Dia merupakan lulusan SMA/Sederajat yang berprofesi sebagai Ibu Rumah Tangga.
Berikut adalah motivasi dan target sasaran dari Ella Rawita sesuai diambil dari
halaman kpu:


Dokumen 2
Motivasi:
-
Target Sasaran:
-
Berdasarkan data KPU, Ella Rawita memiliki status hukum sebagai berikut:
Profil Caleg Pemilu 2019: Yunita Damanik, Caleg DPRD-II Dapil Kota Tebing Tinggi
3 dari PKPI.
Yunita Damanik, menjadi salah satu calon legislatif (caleg) dalam Pemilu 2019
tingkat DPRD-II, 17 April 2019 mendatang.
Dia akan bertarung di daerah pemilihan (dapil) Kota Tebing Tinggi 3.
Berdasarkan info dari laman Komisi Pemilihan Umum (KPU), Caleg Nomor Urut 6 PKPI
ini lahir di Sibual, 13 Juni 1988.
Dia merupakan lulusan SMA/Sederajat yang berprofesi sebagai Guru.
Berikut adalah motivasi dan target sasaran dari Yunita Damanik sesuai diambil
dari halaman kpu:
```

Figure 5. Example of corpus

As seen in Figure 5, the language structure and word order were almost the same between each document. This appeared in such a large number that it was too simple for the model to guess 15% of the masked sentences. There were 22 GB of news data taken from several sources such as the Indonesian Wikipedia, Jakarta News, Okezone, Antara, Kumparan, Tribune, and OSCAR. For language model training it used the RoBERTa (the comparison of performance model language can see on Table 1).

Table 1. Comparison of performance model language

Model language	Exact match (EM)	F1	Accuracy
RoBERTa	84.6%	86.2%	91.7%
AlBERT	82.3%	84.6%	89.7
ELECTRA	83.1%	85.4%	87.3

After conducting training 30% of the total agreed steps, the model showed a fairly good performance where, when using the same dataset from the previous model, there was an increase of about 2% to 4% in the exact match (EM) and F1 matrices. As a result, in the 30% trained model, the model produced 84.6% EM and 86.2% F1 performance. From these results, the model has been able to outperform the ELECTRA model, which was trained using the Malaysian language in the previous progress we used as the baseline performance. This result could also continue to increase until 100% of the model was trained.

4. CONCLUSION

We evaluated the performance of the ALBERT, RoBERTa, ELECTRA models using Indonesian, Malaysian and Esperanto languages. For the best result for building Indonesian QAS, the language model used was RoBERTa that produced 84.6 EM and 86.2% F1 performance. This result could also continue to increase until 100% of the model was trained.

In the implementation of real-world cases, the application of RoBERTa alone is still not sufficient to get the right answer. This is because RoBERTa is trained to find answers from the appropriate context of the questions. Our future work will carry out the search for answers in a context adapted to a very large body of knowledge. The search or retrieval method that we use to get the context that matches the question is Okapi best matching (BM25) – a ranking system that is used to sort the results of the similarity (similarity) to the documents used based on the desired keywords.

ACKNOWLEDGEMENTS

The author would like to thank all the facilities provided by the national research and innovation agency (BRIN) particularly Research Center for Data Science and Informatios and The Informatic Departement, Brawijaya University during the process of making this manuscript.

REFERENCES

[1] A. Bayoudhi, L. H. Belguith, and H. Ghorbel, "Question focus extraction and answer passage retrieval," in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, 2014, pp. 658-665, doi: 10.1109/AICCSA.2014.7073262.

[2] A. Bayoudhi, H. Ghorbel, and L. H. Belguith, "Question answering system for dialogues: A new taxonomy of opinion questions," *Flexible Query Answering Systems*, pp. 67–78, 2013, doi: 10.1007/978-3-642-40769-7_6.

[3] A. B. Abacha and P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," *Information Processing and Management*, vol. 51, no. 5, pp. 570–594, 2015, doi: 10.1016/j.ipm.2015.04.006.

[4] W. Suwarningsih, A. Purwarianti, and I. Supriana, "Semantic roles labeling for reuse in CBR based QA system," *2016 4th International Conference on Information and Communication Technology (ICoICT)*, 2016, pp. 1-5, doi: 10.1109/ICoICT.2016.7571935.

[5] M. Tan, C. D. Santos, B. Xiang, and B. Zhou, "LSTM-based Deep Learning Models for Non-factoid Answer Selection," *ArXiv*, 2015, doi: 10.48550/arXiv.1511.04108.

[6] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *Proc. of the 27th International Conference on Computational Linguistics*, 2018, pp. 1638–1649. [Online]. Available: https://aclanthology.org/C18-1139.pdf

[7] B. Kratzwald and S. Feuerriegel, "Learning from on-line user feedback in neural question answering on the web," *The World Wide Web Conference*, 2019, pp. 906–916, doi: 10.1145/3308558.3313661.

[8] M. R. Akram, C. P. Singhabahu, M. S. M. Saad, P. Deleepa, A. Nugaliyadde, and Y. Mallawarachchi, "Adaptive Artificial Intelligent Q&A Platform," *ArXiv*, 2019, doi: 10.48550/arXiv.1902.02162.

[9] L. Cao, X. Qiu, and X. Huang, "Question answering for machine reading with lexical chain," in *CLEF 2011 Labs and Workshop*, 2011. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.656.7700&rep=rep1&type=pdf

[10] L. Sha, X. Zhang, F. Qian, B. Chang, and Z. Sui, "A multi-view fusion neural network for answer selection," *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018, pp. 5422–5429. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11989/11848

[11] A. C. O. Reddy and K. Madhavi, "Convolutional recurrent neural network with template based representation for complex question answering," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 3, pp. 2710–2718, 2020, doi: 10.11591/ijece.v10i3.pp2710-2718.

[12] X. Zhang, M. H. Chen, and Y. Qin, "NLP-QA Framework Based on LSTM-RNN," in *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, 2018, pp. 307-311, doi: 10.1109/ICDSBA.2018.00065.

[13] S. B. Jadhav, V. R. Udupi, and S. B. Patil, "Convolutional neural networks for leaf image-based plant disease classification," *International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 328–341, 2019, doi: 10.11591/ijai.v8.i4.pp328-341.

[14] T. Shao, X. Kui, P. Zhang, and H. Chen, "Collaborative learning for answer selection in question answering," *IEEE Access*, vol. 7, pp. 7337–7347, 2019 doi: 10.1109/ACCESS.2018.2890102.

[15] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," *International Conference on Learning Representations 2017*, 2017, doi: 10.48550/arXiv.1611.01604.

[16] L. E. -Dor *et al.*, "Active Learning for BERT: An Empirical Study," in *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7949–7962, doi: 10.18653/v1/2020.emnlp-main.638.

[17] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, vol. 1, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[18] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *ArXiv*, 2019, doi : 10.48550/arXiv.1907.11692.

[19] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," *ArXiv*, 2019, doi: 10.48550/arXiv.1909.11942.

[20] I. Akermi, J. Heinecke, and F. Herledan, "Transformer based Natural Language Generation for Question-Answering," in *Proc. of the 13th International Conference on Natural Language Generation*, 2020, pp. 349–359. [Online]. Available: https://aclanthology.org/2020.inlg-1.41.pdf

[21] J. Dai, H. Yan, T. Sun, P. Liu, and X. Qiu, "Does syntax matter? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa," in *Proc. of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1816–1829, doi: 10.18653/v1/2021.naacl-main.146.

[22] A. Wen, M. Y. Elwazir, S. Moon, and J. Fan, "Adapting and evaluating a deep learning language model for clinical why-question answering," *JAMIA Open*, vol. 3, no. 1, pp. 16-20, 2020, doi: 10.1093/jamiaopen/ooz072.

[23] J. S. Sharath and R. Banafsheh, "Question Answering over Knowledge Base using Language Model Embeddings," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206698.

[24] W. Antoun, F. Baly, and H. Hajj, "AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding," in *Proc. of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 191-195. [Online]. Available: https://aclanthology.org/2021.wanlp-1.20.pdf

[25] K. Clark, M. -T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," *International Conference on Learning Representations (ICLR) 2020*, 2020, doi: 10.48550/arXiv.2003.10555.

[26] A. Kavros and Y. Tzitzikas, "SoundexGR: An algorithm for phonetic matching for the Greek language," *Natural Language Engineering*, 2022, doi: 10.1017/S1351324922000018.

[27] W. Suwarningsih, R. A. Pratama, F. Y. Rahadika, and M. H. A. Purnomo, "Self-Attention Mechanism of RoBERTa to Improve QAS for e-health Education," *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, 2021, pp. 221-225, doi: 10.1109/IC2IE53219.2021.9649363.

BIOGRAPHIES OF AUTHORS

Wiwin Suwarningsih she was graduated from her bachelor degree at Informatics Program, Adityawarman Institute of Technology Bandung in 1996. She got graduated from her master education in 2000 at the Informatics Study Program, Bandung Institute of Technology and doctoral degree in 2017 the School of Electrical and Informatics Engineering, Bandung Institute of Technology. Since 2006 until now she has been a researcher at Research Center for Information and Data Science, National Research and Innovation Agency, Indonesia. His research interests are artificial intelligence, computational linguistics, particularly in Indonesian natural language processing and Indonesian text mining, information retrieval, and question answering systems. She can be contacted at email: wiwin.suwarningsih@brin.go.id.

Raka Aditya Pratama currently He is still studying at Informatics department, Brawijaya University as a bachelor student. His research interests are on Image processing, Artificial Inteligence, Phyton programming, Text mining, and Software Engineering. He can be contacted at email: rakdit@student.ub.ac.id.

Fadhil Yusuf Rahadika currently He is still studying at Informatics department, Brawijaya University as a bachelor student. His research interests are on Artificial Inteligence, Data mining, Software engineering and Phyton programming. He can be contacted at email: fadhilyusuf27@student.ub.ac.id.

Mochamad Havid Albar Purnomo currently He is still studying at Informatics department, Brawijaya University as a bachelor student. His research interests are on Text mining, Phyton programming, Software engineering and Artificial Inteligence. He can be contacted at email: havidalbar@student.ub.ac.id.