Machine Reading Comprehension untuk Teks Bahasa Indonesia

Zendy Bramantia Alfareza[1], Ade Romadhony[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]zendybramantia@student.telkomuniversity.ac.id, [2]aderomadhony@telkomuniversity.ac.id

Abstrak

Membaca merupakan suatu keterampilan yang sangat penting dan harus dikuasai oleh semua orang. Metode yang sering digunakan dalam meningkatkan kemampuan pemahaman membaca adalah pelajaran pemahaman membaca (reading comprehension). Dalam mempermudah pembuatan *assessment reading comprehension* terdapat *task* pada bidang *Natural Language Processing* (NLP) yakni *Machine Reading Comprehension*(MRC). MRC adalah *task* dasar dari *question answering*(QA), di mana pada setiap pertanyaan diberikan konteks terkait untuk memprediksi jawabannya. Tujuan MRC adalah untuk memprediksi jawaban yang benar dari konteks yang diberikan atau bahkan menghasilkan jawaban yang lebih kompleks berdasarkan konteks yang diberikan. Model yang digunakan merupakan model BERT yang sudah dilatih untuk memahami teks bahasa Indonesia atau bisa disebut dengan IndoBERT. Penelitian ini menggunakan dataset berupa 99 soal reading comprehension. Hasil pengujian menunjukkan bahwa pada learning rate 1e-5, prediksi jawaban pada tugas pemahaman baca ini mencapai kinerja terbaik dengan akurasi sebesar 75% dan skor F1 sebesar 92%.

Kata kunci : BERT, IndoBERT, *Reading Comprehension*, QA, NLP

Abstract

Reading is a crucial skill that is essential for everyone to master. A commonly used method to enhance reading comprehension abilities is through reading comprehension lessons. To facilitate the creation of reading comprehension assessments, there exists a task in the field of Natural Language Processing (NLP) known as machine reading comprehension (MRC). machine reading comprehension serves as a fundamental task in question answering (QA), where each question is provided with relevant context to predict its answer. The goal of machine reading comprehension is to predict the correct answers from the given context, and in some cases, even generate more complex answers based on the provided context. The model employed in this study is the BERT model, which has been pre-trained to understand Indonesian language texts, also known as IndoBERT. This research employs a dataset consisting of 99 reading comprehension questions. The testing results demonstrate that using a learning rate of 1e-5, the answer predictions in this reading comprehension task achieve the best performance with an accuracy of 75% and an F1 score of 92%.

Keywords: BERT, IndoBERT, Reading Comprehension, QA, NLP

1. Pendahuluan

Latar Belakang

Membaca merupakan suatu keterampilan yang sangat penting dan harus dikuasai oleh semua orang. Namun bisa membaca saja tidak cukup, pembaca juga harus memahami maksud dari teks yang dibaca. Metode yang sering digunakan dalam meningkatkan kemampuan pemahaman membaca adalah pelajaran pemahaman membaca (*reading comprehension*). *Reading comprehension* merupakan pilar utama selama kegiatan membaca untuk membangun pemahaman teks. Meningkatkan keterampilan *reading comprehension* dapat berdampak positif pada banyak aspek kinerja akademik siswa. Cara kerja dari *reading comprehension* adalah dengan diberikan *assessment* berupa konteks, pertanyaan dan pilihan jawaban [1]. Lalu siswa memilih jawaban yang benar berdasarkan teks dan pertanyaan yang diberikan. Dalam mempermudah pembuatan *assessment reading comprehension* terdapat task pada bidang *Natural Language Processing* (NLP) yakni *Machine Reading Comprehension.*

Machine reading comprehension adalah task dasar dari question answering(QA), di mana pada setiap pertanyaan diberikan konteks terkait untuk memprediksi jawabannya[2]. Tujuan dari machine reading comprehension adalah untuk membantu dalam pembuatan *assessment reading comprehension*. Dalam kasus ini *Bidirectional Encoder Representations from Transformers* (BERT) dapat mengenali pertanyaan yang berkaitan dengan konteks dan mendapatkan bagian yang paling relevan dari konteks sebagai jawaban atas pertanyaan tersebut. BERT didesain untuk memahami teks yang tidak memiliki label terlebih dahulu dengan memperhatikan kata-kata di sebelah kiri dan kanan dalam setiap lapisannya. Ini membuat BERT bisa belajar tentang konteks dari

kedua arah [3].Maka dari itu pada penelitian ini model yang digunakan adalah BERT.

Penelitian ini bertujuan untuk memprediksi jawaban soal Reading Comprehension bahasa Indonesia. Dataset pada penelitian ini menggunakan bahasa indonesia, maka model yang digunakan merupakan model BERT yang sudah dilatih untuk memahami teks bahasa Indonesia atau bisa disebut dengan *IndoBERT*[4].

1.1 Topik Dan Batasan

Penelitian ini mengangkat topik sistem identifikasi jawaban pada kumpulan pertanyaan pilihan ganda. **s**

1.2 Tujuan

Tujuan dari penelitian ini adalah mengimplementasikan sistem identifikasi jawaban pada kumpulan pertanyaan pilihan ganda.

2. Studi Terkait

Pada penelitian sebelumnya, Seo M beserta rekan-rekannya melakukan penelitian untuk *task* MRC dengan menggunakan model *Bi-Directional Attention Flow model* (BiDAF) dan menggunakan algoritma CNN. Penelitian tersebut menggunakan dataset the SQuAD yang dirilis oleh Stanford NLP Group dan terdiri dari 100,000 pasangan jawaban dan paragraf konteks untuk setiap pasangan. Penelitian ini berhasil mencapai EM sebesar 73,3% dan F1 sebesar 81,1% [5].

Park C beserta rekan-rekannya melakukan penelitian MRC menggunakan model S2-Net dan simple recurrent unit (SRU), dan menggunakan algoritma RNN. Penelitian tersebut menggunakan dataset bahasa Korea. Dataset yang digunakan berupa pasangan dari sebuah paragraf, pertanyaan dan jawaban yang didapatkan dari berita dan Wikipedia untuk hiburan, dan menggunakan format yang mirip dengan dataset the SQuAD. Hasil eksperimen menunjukkan bahwa S2-Net menunjukkan kinerja terbaik dengan dataset pemahaman mesin berbahasa Korea, dengan hasil 68,95% EM dan 81,15% F1 [6].

Kegang Xu beserta rekan-rekannya melakukan penelitian untuk memprediksi jawaban pada soal reading comprehension untuk pilihan ganda menggunakan model BERT. Dataset yang digunakan merupakan dataset RACE. Penelitian ini berhasil mencapai akurasi sebesar 66,2% dalam model tunggal dan 67,9% dalam model gabungan[7]

2.1 Klasifikasi Teks

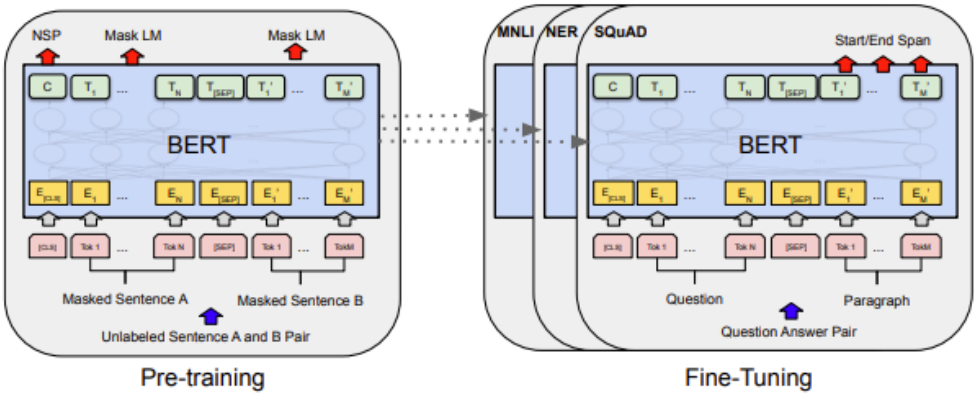
Klasifikasi teks adalah proses pengelompokkan dokumen-dokumen teks ke dalam berbagai kategori. Bentuk paling umumnya adalah klasifikasi biner, di mana setiap dokumen ditempatkan ke dalam salah satu dari dua kategori. Dalam pendekatan yang lebih terstruktur, klasifikasi teks adalah proses analisis yang mengambil dokumen teks apa pun sebagai masukan dan memberinya klasifikasi dari sejumlah label kelas yang sudah ditentukan sebelumnya[8].

2.2 Machine Reading Comprehension

Machine Reading Comprehension (MRC) adalah task yang diperkenalkan untuk menguji sejauh mana mesin dapat memahami bahasa alami dengan meminta mesin untuk menjawab pertanyaan berdasarkan konteks tertentu, yang berpotensi merevolusi cara manusia dan mesin berinteraksi dengan satu sama lain [9]. MRC adalah task dasar dari *question answering*(QA), di mana pada setiap pertanyaan diberikan konteks terkait untuk memprediksi jawabannya. Tujuan MRC adalah untuk memprediksi jawaban yang benar dari konteks yang diberikan atau bahkan menghasilkan jawaban yang lebih kompleks berdasarkan konteks yang diberikan [5].

2.3 Bidirectional Encoder Representations from Transformers

BERT adalah model NLP yang dirancang untuk memahami teks yang tidak berlabel dan kemudian disesuaikan lebih lanjut menggunakan teks yang sudah diberi label untuk berbagai tugas NLP. Dengan BERT, kita bisa membuat model-model NLP terkini untuk berbagai tugas, seperti menganalisis teks atau mengerti bahasa manusia. Model BERT yang sudah dilatih sebelumnya dapat dilakukan *Fine-tuning* dengan menambahkan hanya satu lapisan output tambahan untuk membuat model-model canggih yang cocok untuk banyak *task*, seperti menjawab pertanyaan dan menyimpulkan bahasa, tanpa perlu mengubah struktur inti model secara besar-besaran untuk setiap tugas yang lebih spesifik. Pada model BERT, untuk melakukan pre-training dan fine tuning menggunakan arsitektur yang sama. Dilakukan encoding untuk setiap input menggunakan token spesial [CLS] yang merupakan token klasifikasi, dan [SEP] yang merupakan token *separator* [3]. Untuk detil lebih jelas dapat dilihat pada gambar 1.



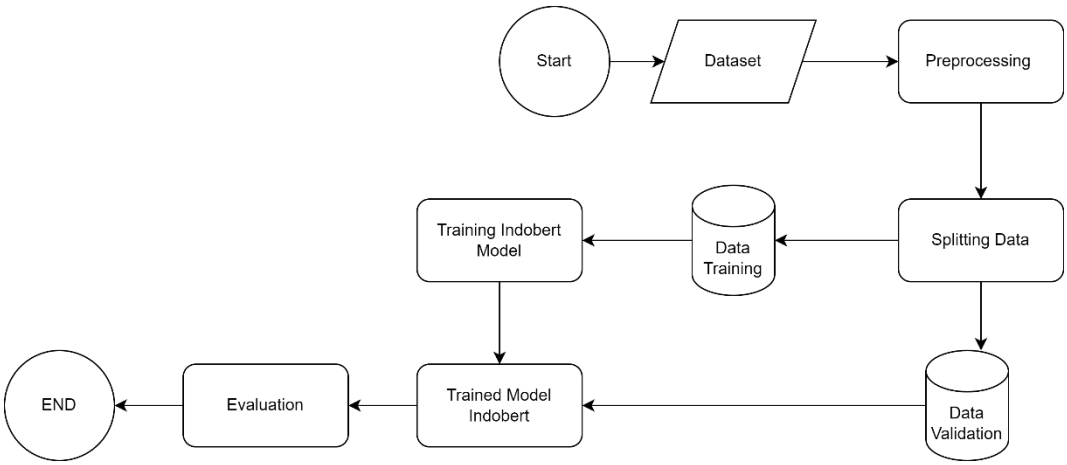
Gambar 1 bert Pre-training dan Fine-tuning[3]

2.4 IndoBERT

IndoBERT adalah sebuah model berbasis transformer yang dikembangkan dengan mengikuti gaya model BERT, tetapi difokuskan pada bahasa Indonesia. Model ini dilatih sebagai sebuah masked language model menggunakan *framework Huggingface* [4]. *IndoBERT* dilatih menggunakan dataset yang terdiri dari sekitar 4 miliar korpus kata(Indo4B) dan didapatkan dari sumber yang tersedia untuk umum seperti sosial media, blog, berita, dan situs web.[10]

3. Sistem yang Dibangun

Tujuan dari penelitian ini adalah untuk mengimplementasikan *Task Machine Reading Comprehension* untuk memprediksi jawaban dari konteks dan pertanyaan yang diberikan. Penelitian ini memiliki beberapa tahapan yaitu pengumpulan dataset, preprocessing dataset, prediksi jawaban menggunakan BERT, dan Evaluasi Model, seperti yang terlihat pada Gambar 2.



Gambar 2. Flowchart Perancangan Sistem

3.1 Dataset

Pada penelitian ini dataset yang digunakan berasal dari penelitian yang dilakukan oleh Daryannor et al yang berupa 100 soal reading comprehension pilihan ganda yang telah diterjemahkan ke bahasa Indonesia[11]. Penelitian tersebut menggunakan dataset MCTest berjenis MC160, yang bersumber dari penelitian Richardson et al[12]. Pada dataset MC160 terbagi menjadi 3 bagian, yakni train, dev dan test. Namun pada penelitian ini, bagian dataset yang digunakan hanya sebesar 44 data dev dan 55 data train. Untuk gambaran dataset yang digunakan dapat dilihat pada tabel 1.

Tabel 1. Dataset

Nama Kolom	Deskripsi
Paragraf Konteks	Bagian teks yang menyediakan informasi dan detail yang relevan untuk menjawab pertanyaan yang diajukan.
Pertanyaan	pernyataan tertulis yang diajukan berdasarkan isi pada paragraf konteks.
Pilihan Jawaban A	opsi yang disediakan sebagai potensi jawaban untuk pertanyaan yang diajukan terkait dengan teks yang telah dibaca

Pilihan Jawaban B	opsi yang disediakan sebagai potensi jawaban untuk pertanyaan yang diajukan terkait dengan teks yang telah dibaca
Pilihan Jawaban C	opsi yang disediakan sebagai potensi jawaban untuk pertanyaan yang diajukan terkait dengan teks yang telah dibaca
Pilihan Jawaban D	opsi yang disediakan sebagai potensi jawaban untuk pertanyaan yang diajukan terkait dengan teks yang telah dibaca
Label	Indeks jawaban yang benar atau yang diharapkan dari pertanyaan yang diajukan terhadap teks yang telah diberikan

3.2 Splitting Data

Pada penelitian ini dilakukan *splitting data* menjadi 2 bagian, yakni data latih dan data validasi. Data latih digunakan untuk melatih model dan data validasi digunakan untuk validasi model. Pada tahap *splitting data* ini 80% data akan digunakan untuk melatih model *IndoBERT* dan 20% data akan digunakan untuk melakukan evaluasi model IndoBERT yang sudah dilatih sebelumnya. Pada penelitian ini tidak menggunakan data test dikarenakan dataset yang digunakan hanya sebesar 99 data.

Tabel 2. Pembagian Dataset

Label	Data Latih	Data Validasi
Soal Reading Comprehension	79	20

3.3 Preprocessing

Setelah memiliki data set untuk digunakan, langkah selanjutnya adalah pemrosesan awal atau *preprocessing*. Pada preprocessing ini akan dilakukan penggabungan kalimat dengan menggunakan *special token*. Token yang akan digunakan yaitu CLS, yakni token yang mengawali teks yang akan dimasukan, SEP, yang digunakan untuk memisahkan 2 teks, dan PAD yakni token yang digunakan untuk *padding*. Pada penelitian akan diberikan 4 masukan untuk setiap pertanyaan.

Input :
[CLS] *context* [SEP] *question* + *option* 1 [SEP]
[CLS] *context* [SEP] *question* + *option* 2 [SEP]
[CLS] *context* [SEP] *question* + *option* 3 [SEP]
[CLS] *context* [SEP] *question* + *option* 4 [SEP]

3.4 Fine Tuning Model

Pada tahap fine tuning model terdapat beberapa parameter untuk model BERT. Pada parameter *learning rate* memiliki nilai 1e-5 karena nilai tersebut mendapatkan hasil evaluasi yang cukup tinggi, lalu terdapat parameter *batch size* yang memiliki nilai 8, karena jika *batch size* memiliki nilai yang lebih tinggi terjadi *memory leak* dan proses pelatihan akan terhenti, parameter *epoch* memiliki nilai 20 karena pada saat pelatihan hasil *evaluation loss* mencapai hasil terbaik di sekitar epoch ke 15-25, parameter *weight decay* dengan nilai 0.01 yang merupakan nilai parameter *default*, parameter *metric for best model* yakni ‘eval loss’ yang berfungsi untuk memberhentikan pelatihan ketika *evaluation loss* telah berada pada nilai paling rendah dan parameter *load best model at end* dengan nilai *true* yang digunakan untuk memilih model dengan hasil *eval loss* terbaik saat pelatihan berhenti.

Tabel 3. Parameter

Parameter	**Value**
learning rate	1e-5
batch size	8
epoch	20
weight decay	0.01
metric for best model	‘eval loss’
load best model at end	*true*

3.5 Evaluasi Model

Evaluasi adalah tahap yang penting untuk mengetahui performansi pada suatu model atau metode yang digunakan sebuah penelitian. Penelitian ini menggunakan *F1-Score* untuk mengevaluasi performa model pada task machine reading comprehension untuk teks bahasa Indonesia. Dalam menghitung *F1-Score* dibutuhkan tabel confusion matrix untuk menghitung nilai precision dan recall[13].

Pada penelitian ini tahap evaluasi dilakukan dengan menggunakan *confusion matrix* (CM) dengan matriks utama yaitu Accuracy dan F1-Score. Terdapat 4 (empat) istilah yang digunakan untuk menjelaskan hasil proses pemodelan, yakni *True Positive* (TP) merupakan jumlah data positif yang diprediksi benar, *True Negative* (TN) merupakan jumlah data negatif yang diprediksi benar, *False Positive* (FP) merupakan jumlah data negatif namun diprediksi sebagai data positif, dan *False Negative* (FN) merupakan jumlah data positif namun diprediksi sebagai data negatif. Contoh penggunaan confusion matriks dituliskan pada tabel 2. Pada eksperimen ini TP didapatkan dari jumlah jawaban benar yang berhasil diprediksi dengan benar, FP didapatkan dari jumlah jawaban salah yang diprediksi dengan benar, TN didapatkan dari jumlah jawaban salah yang diprediksi dengan salah, FN didapatkan dari jumlah jawaban benar yang diprediksi dengan salah.

Tabel 4. Confusion Matrix

Confusion Matrix		*Actual Value*	
		Positif	Negatif
Prediction Value	Positif	TP	FP
	Negatif	FN	TN

Setelah mendapatkan nilai TP, TN, FP, dan FN dapat diperoleh nilai *accuracy, precision, recall,* dan *f1-score* dengan persamaan seperti pada tabel berikut:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4. Evaluasi

4.1 Hasil Pengujian

Penelitian ini melibatkan penggunaan 79 data latih yang digunakan dalam pelatihan model *IndoBERT*. Model ini dikembangkan dengan menerapkan 5 skenario yang berbeda, dan setiap skenario memiliki *learning rate* yang berbeda pula. Detail skenario tersebut akan dipaparkan dalam Tabel 5.

Penelitian ini menggunakan 5 skenario dengan nilai *learning rate* yang berbeda untuk mendapatkan hasil evaluasi model yang terbaik. Setelah dilakukan pengujian dengan menggunakan berbagai skenario sebelumnya, hasilnya terlihat pada Tabel 6 bahwa skenario kedua menunjukkan nilai *F1-Score* tertinggi dibandingkan dengan skenario lainnya. Pada skenario kedua, tercapai akurasi sebesar 75% dan *F1-Score* sebesar 92%, yang menandakan bahwa skenario ini memiliki performa terbaik dibandingkan dengan skenario lainnya. Learning rate

Tabel 5. Skenario

Skenario	**Learning Rate**
1	9e-6
2	**1e-5**
3	2e-5
4	3e-5
5	5e-5

Tabel 6. Hasil pengujian

Skenario	**Accuracy**	**Precision**	**Recall**	**F1-Score**
1	70%	75%	100%	85%
2	**75%**	**85%**	**100%**	**92%**
3	65%	71%	83%	76%
4	60%	71%	83%	76%
5	55%	80%	66%	72%

4.2 Analisis Hasil Pengujian

Setelah dilakukan pelatihan pada model *IndoBERT*, terlihat bahwa skenario ke-2 mendapatkan hasil pengujian yang cukup baik dengan nilai akurasi 75%, nilai *precision* 85% , nilai *recall* 100% dan *F1-Score* 92%. Hasil tersebut menunjukkan bahwa 79 data latih dapat digunakan untuk melatih model IndoBERT dan mendapatkan hasil evaluasi yang cukup baik.

Setelah dilakukan pelatihan pada model *IndoBERT*, ditemukan bahwa terdapat 15 data yang berhasil dijawab dengan benar, sementara 5 data tidak berhasil dijawab. Data yang berhasil dijawab dengan benar umumnya memiliki kaitan dengan konteks pertanyaan, begitu pula sebaliknya. Contoh data yang berhasil dijawab dengan benar dan data yang tidak berhasil dijawab dengan benar dapat dilihat pada Tabel 7 dan 8.

Tabel 7. Contoh 2 data dari 15 data validasi yang berhasil menjawab soal dengan benar.

Soal	Konteks	Pertanyaan	Label	Prediksi
1	Suatu ketika ada seorang putri yang tinggal di suatu menara tinggi...	Dimanakah putri tersebut tinggal pertama kali ?	Menara tinggi	Menara tinggi
2	Peter adalah seekor anak anjing yang merasakan kesedihan…	Anak anjing yang seperti apa yang diinginkan Sammie ?	Anak anjing emas yang dapat berbaring dengannya	Anak anjing emas yang dapat berbaring dengannya

Dapat dilihat pada tabel 7 dengan pertanyaan Dimanakah putri tersebut tinggal pertama kali ?" yang mempunyai label "Menara tinggi" terdapat kalimat atau informasi yang sesuai pada konteks " Suatu ketika ada seorang putri yang tinggal di suatu menara tinggi...", dengan informasi "putris tersebut tinggal di suatu menara tinggi" tersebut berkaitan dengan pertanyaan pada konteks tersebut, sehingga pertanyaan tersebut bisa terjawab.

Tabel 8. Contoh 2 data dari 5 data validasi yang tidak berhasil menjawab soal dengan benar.

Soal	Konteks	Pertanyaan	Label	Prediksi
1	Suatu ketika ada tiga kelinci yang bernama Winston...	Apa hewan peliharan yang sedang bermain dengan Johnny di halaman ?	Chester	Winston
2	Sally suka pergi keluar. Dia memakai sepatu...	Siapa nama kucing itu ?	Missy	Meow

Sedangkan pada contoh data pada soal nomor 1 tabel 5, pertanyaan yang diajukan adalah "Apa hewan peliharan yang sedang bermain dengan Johnny di halaman?" Namun, dalam konteks yang diberikan, tidak ada informasi yang mengacu pada karakter "Johnny" atau aktivitas mereka bermain di halaman. Semua informasi dalam teks berkaitan dengan tiga kelinci bernama Winston, Chester, dan Francis, serta interaksi mereka dengan taman wortel. Karena tidak ada informasi yang merujuk kepada "Johnny" atau aktivitas di halaman, tidak mungkin menjawab pertanyaan tersebut dengan benar berdasarkan konteks yang diberikan dalam teks.

Dan pada contoh data pada soal nomor 2 tabel 5 hasil prediksi dari model *IndoBERT* menunjukkan bahwa pilihan jawaban yang diprediksi sebagai jawaban yang benar adalah "Meow", sementara jawaban yang seharusnya benar adalah "Missy". Meskipun pertanyaan tentang nama kucing cukup sederhana, konteks teks yang lebih luas mungkin membuat model bingung. Model mungkin tidak dapat dengan tepat mengenali bahwa "Missy" adalah nama kucing yang disebutkan dalam konteks.

5. Kesimpulan

Dalam tugas akhir ini, dilakukan prediksi jawaban untuk soal *Reading Comprehension* menggunakan bahasa Indonesia. Model *IndoBERT* digunakan untuk melakukan prediksi jawaban. Selain itu, juga dilakukan perbandingan performansi model dengan variasi *learning rate* yang berbeda. Learning rate yang digunakan antara lain 9e-6, 1e-5, 2e-5, 3e-5 dan 5e-5.

Pada penelitian ini, dataset berisi 99 soal pemahaman baca dalam bahasa Indonesia digunakan. Pengujian dilakukan dengan memanfaatkan matriks kebingungan (confusion matrix) untuk menghitung nilai akurasi dan skor F1. Hasil pengujian menunjukkan bahwa pada learning rate 1e-5, prediksi jawaban pada tugas pemahaman baca ini mencapai kinerja terbaik dengan akurasi sebesar 75% dan skor F1 sebesar 92%.

Berdasarkan analisis hasil pengujian, terdapat dugaan bahwa hasil prediksi yang tidak tepat mungkin disebabkan oleh kompleksitas konteks dan ataupun kurangnya hubungan antara pertanyaan dengan konteks paragraf atau jawaban yang disediakan. Oleh karena itu, model *IndoBERT* mengalami keterbatasan dalam kemampuan prediksi yang optimal.

Daftar Pustaka

[1] N. Fitri and Y. Zainil, "Journal of English Language Teaching Enhancing College Students' Reading Comprehension Through Critical Reading," *Journal of English Language Teaching*, vol. 7, no. 4, [Online]. Available: http://ejournal.unp.ac.id/index.php/jelt

[2] I. Nguyen and H. Thinh, "RNN on Machine Reading Comprehension-Bi-Directional Attention Flow model."

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[4] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.00677

[5] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional Attention Flow for Machine Comprehension," Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.01603

[6] C. Park *et al.*, "S2-Net: Machine reading comprehension with SRU-based self-matching networks," *ETRI Journal*, vol. 41, no. 3, pp. 371–382, Jun. 2019, doi: 10.4218/etrij.2017-0279.

[7] "A_BERT_based_model_MCRC".

[8] "Text Classification and Categorization," in *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Elsevier Inc., 2012, pp. 881–892. doi: 10.1016/B978-0-12-386979-1.00035-9.

[9] S. Liu, X. Zhang, S. Zhang, H. Wang, and W. Zhang, "Neural machine reading comprehension: Methods and trends," *Applied Sciences (Switzerland)*, vol. 9, no. 18. MDPI AG, Sep. 01, 2019. doi: 10.3390/app9183698.

[10] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Sep. 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[11] D. Ibnu, "Eksplorasi Reading Comprehension Berbasis Open Information Extraction Bahasa Indonesia.", 2020

[12] M. Richardson, C. J. C. Burges, and E. Renshaw, "MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text," 2013. [Online]. Available: http://www.mturk.com

[13] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, Jan. 2020, doi: 10.1186/s12864-019-6413-7.