IMPLEMENTASI TRANSFER LEARNING DENGAN INDOBERT PADA QUESTION ANSWERING BAHASA INDONESIA

Arniyantika Putri Mediani, Farid Akbar Siregar
Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara
alamat umsu utamaJl. Kapt. Mukhtar Basri No. 3 Medan, 20238, Sumatera Utara.
arniyantika.putriii@gmail.com

ABSTRAK

Penelitian ini membahas pengembangan sistem Question Answering (QA) berbasis teks yang mampu menjawab pertanyaan secara otomatis dengan mengandalkan konteks yang diberikan oleh pengguna. Sistem dirancang menggunakan model pretrained BERT (Bidirectional Encoder Representations from Transformers), yang mampu memahami hubungan antara pertanyaan dan konteks untuk menentukan jawaban secara tepat. Proses kerja sistem meliputi tokenisasi, penerapan multi-head self-attention melalui arsitektur Transformer, serta perhitungan nilai start logit dan end logit untuk menentukan rentang teks yang paling mungkin sebagai jawaban. Sistem tidak hanya menampilkan satu jawaban terbaik, tetapi juga menyajikan sejumlah kemungkinan jawaban lain yang disusun berdasarkan skor akurasi. Evaluasi dilakukan menggunakan metrik Exact Match (EM) dan F1 Score untuk mengukur kesesuaian antara jawaban sistem dan jawaban referensi. Hasil penelitian menunjukkan bahwa pendekatan ini efektif dalam mengidentifikasi jawaban yang relevan dari konteks panjang dengan efisiensi dan tingkat ketepatan yang tinggi. Sistem ini berpotensi diterapkan dalam berbagai kebutuhan pencarian informasi otomatis, seperti layanan bantuan digital, chatbot, dan aplikasi pendidikan.

Kata kunci: *Question Answering, BERT, Start–End Logit, F1 Score, Multi-head Attention*

1. PENDAHULUAN

Salah satu fokus penting dalam pemrosesan bahasa alami (NLP) adalah pengembangan sistem *Question Answering* (QA) dalam Bahasa Indonesia. Sistem *Question Answering* merupakan suatu sistem yang didalamnya menggunakan *Natural Language Processing* (NLP) sebagai pemrosesan bahasa alami manusia untuk menggali informasi kemudian secara otomatis menjawab pertanyaan yang telah diajukan [1]. Kemajuan dalam teknologi *transfer learning*, khususnya dengan model besar seperti *Bidirectional Encoder Representations for Transformers* (BERT), telah membawa perubahan signifikan dalam pendekatan pemrosesan bahasa alami. *IndoBERT*, sebagai model BERT yang dirancang khusus untuk Bahasa Indonesia, dilatih menggunakan korpus teks besar dari berbagai domain, memungkinkan model ini memahami konteks bahasa dengan lebih baik dibandingkan model generik seperti BERT multibahasa [2]. Dengan kemajuan yang cepat dari teknologi dan informasi pada saat ini, hingga kita semakin sulit untuk terlepas dari perkembangannya. Hal ini menyebabkan arus informasi berkembang dengan cepat [3]. Pengiriman informasi saat ini dapat dilakukan dengan cepat dan mudah, salah satunya melalui aplikasi tanya jawab atau Question Answering System (QAS) terkait materi yang ingin diketahui oleh pengguna [4]. QAS adalah suatu sistem yang memungkinkan pengguna untuk mengajukan pertanyaan dalam bahasa sehari-hari dan mendapatkan jawaban secara cepat dan ringkas, kadang-kadang disertai dengan informasi tambahan untuk mendukung kebenaran jawaban tersebut [5]. Dalam konteks analisis sentimen teks Bahasa Indonesia, penelitian terkait penerapan metode transfer learning juga pernah dilakukan sebelumnya. Dalam penelitian tersebut ditemukan bahwa model yang telah dilakukan proses pre-train seperti Vader, Textblob, dan Flair memiliki performa yang lebih unggul dibandingkan dengan model klasifikasi tradisional dalam kasus analisis sentimen [6]. Selain itu, sebelumnya juga pernah dilakukan beberapa penelitian terkait penerapan metode transfer learning pada model IndoBERT. Sebuah penelitian membahas tentang implementasi model IndoBERT dalam analisis sentimen untuk dataset review aplikasi mobile di Indonesia. Temuan penelitian menunjukkan bahwa model IndoBERT mencapai rata-rata akurasi yang lebih tinggi dalam memproses data berbasis leksikon dibandingkan dengan model BERT [7]. Pada penelitian lain, penerapan metode transfer learning pada model IndoBERT untuk analisis sentimen pada data review aplikasi kesehatan juga pernah dilakukan. Hasil penelitian menunjukkan bahwa penerapan metode transfer learning dengan model IndoBERT pada data kesehatan menghasilkan tingkat akurasi yang tinggi, mencapai 96% akurasi, 95% presisi, dan 95% F1-score [8]. Metode transfer learning merupakan teknik pembelajaran mesin yang memanfaatkan pengetahuan yang telah dipelajari dari model sebelumnya untuk menyelesaikan tugas baru [9]. Maka dari itu, dengan menerapkan metode transfer learning model ini memberikan peluang baru untuk mengatasi tantangan QA dalam Bahasa Indonesia. Dalam penelitian ini, dataset IndoLEM (*Indonesian Language Evaluation and Benchmark*) digunakan sebagai sumber data utama. IndoLEM menyediakan data terstruktur untuk tugas-tugas NLP, termasuk QA, yang dirancang untuk mengevaluasi kemampuan model memahami teks dan menjawab pertanyaan dalam Bahasa Indonesia.

Dataset ini sangat relevan untuk membangun dan mengevaluasi sistem QA berbasis IndoBERT. Penelitian ini membantu dalam membangun sebuah aplikasi QA berbahasa Indonesia dengan memanfaatkan model transfer learning IndoBERT dan dataset IndoLEM. Aplikasi ini dirancang untuk memiliki antarmuka yang *user-friendly* dan mampu secara otomatis menjawab pertanyaan berbasis teks dengan akurat. Untuk memastikan kualitas jawaban yang dihasilkan, kinerja model dievaluasi menggunakan metrik seperti skor *Exact Match* (EM) dan F1. Dengan menggabungkan metode IndoBERT dan dataset IndoLEM, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem pemrosesan bahasa alami yang praktis, akurat, dan dapat diterapkan. Dataset ini sangat relevan untuk membangun dan mengevaluasi sistem QA berbasis IndoBERT. Penelitian ini membantu dalam membangun sebuah aplikasi QA berbahasa Indonesia dengan memanfaatkan model transfer learning IndoBERT dan dataset IndoLEM. Aplikasi ini dirancang untuk memiliki antarmuka yang *user-friendly* dan mampu secara otomatis menjawab pertanyaan berbasis teks dengan akurat. Untuk memastikan kualitas jawaban yang dihasilkan, kinerja model dievaluasi menggunakan metrik seperti skor *Exact Match* (EM) dan F1. Dengan menggabungkan metode IndoBERT dan dataset IndoLEM, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem pemrosesan bahasa alami yang praktis, akurat, dan dapat diterapkan.

2. TINJAUAN PUSTAKA.

Penelitian *Intermediate-task Transfer Learning* menunjukkan bahwa penggunaan tugas antara (*intermediate task*) dapat memperbaiki kinerja model NLP dalam konteks Indonesia [19]. melaporkan bahwa penerapan *transfer learning* pada model BERT (termasuk IndoBERT) untuk *closed-domain QA* di situs pendidikan menghasilkan peningkatan F1 hingga 4,91 % dibandingkan tanpa transfer learning [20]. semua tugas pemahaman bahasa alami (NLU) dapat dipetakan menjadi tugas *question answering*, yang mendukung paradigma transfer learning dalam model Bahasa [22]. Dalam penelitian domain spesifik, Ramadhan et al. menunjukkan bahwa IndoBERT-QA mampu mencapai nilai *Mean Reciprocal Rank (MRR)* sebesar 0,91 pada konteks domain tertentu [23]. Pandya et al. mengusulkan teknik *cascading adaptors* untuk mentransfer data dari bahasa Inggris ke bahasa dengan sumber daya terbatas dalam sistem QA. [24]. Raffel et al. dalam studi mereka mengenai transfer learning menyajikan kerangka *text-to-text* yang menjadikan berbagai tugas NLP dapat dipandang sebagai soal *question answering*. [25].

2.1. Pengenalan NLP dan Transfer Learning

Natural Language Processing (NLP) merupakan cabang ilmu komputer dan kecerdasan buatan (Artificial Intelligence/AI) yang berfokus pada bagaimana komputer dapat memahami, memproses, dan berkomunikasi menggunakan bahasa manusia. NLP memanfaatkan kombinasi teknik linguistik, aturan bahasa, statistik, serta metode pembelajaran mesin untuk mengenali dan menghasilkan teks atau ucapan. Perkembangan NLP telah memberikan kontribusi besar terhadap teknologi modern, seperti chatbot, mesin pencari, asisten digital (Siri, Alexa, Cortana), sistem navigasi suara, hingga analisis teks dalam bisnis untuk meningkatkan efisiensi dan produktivitas[10]. Dalam konteks pembelajaran mesin, transfer learning menjadi teknik penting karena memungkinkan model yang telah dilatih pada data besar dan tugas umum digunakan kembali untuk menyelesaikan tugas lain yang lebih spesifik dengan dataset yang lebih kecil. Dalam NLP, metode ini memanfaatkan model bahasa besar yang telah dipra-latih seperti BERT dan GPT untuk meningkatkan kinerja pada tugas seperti analisis sentimen, penerjemahan mesin, atau question answering dengan waktu pelatihan yang lebih efisien.

2.2. Pengolahan Awal Teks (Preprocessing Text)

Pengolahan awal teks atau preprocessing merupakan tahap krusial dalam pemrosesan bahasa alami karena bertujuan untuk membersihkan dan mempersiapkan data teks agar dapat dianalisis secara efektif. Langkah pertama adalah menghapus data duplikat (remove duplicate) untuk mencegah bias interpretasi data [5]. Selanjutnya, dilakukan case folding untuk menyeragamkan huruf menjadi huruf kecil sehingga format teks menjadi konsisten . Proses berikutnya adalah penghapusan tanda baca, angka, dan simbol agar teks menjadi lebih bersih. Setelah itu dilakukan tokenisasi, yaitu proses memecah teks menjadi unit-unit kata atau token yang memudahkan analisis. Tahapan lain adalah menghapus stopwords, yaitu kata-kata umum yang tidak memberikan makna penting dalam konteks analisis [11]. Terakhir, dilakukan stemming untuk mengembalikan kata ke bentuk dasarnya, sehingga model dapat memahami makna inti dari teks tersebut.

2.3. Visualisasi

Visualisasi data merupakan teknik untuk menyajikan informasi secara grafis guna memudahkan pemahaman pola dan hubungan dalam data. Melalui visualisasi, data yang kompleks dapat ditampilkan dalam bentuk interaktif yang memungkinkan pengguna melakukan pengamatan dan analisis secara lebih mendalam [12]. Selain itu, visualisasi juga memfasilitasi penemuan ilmiah dengan mengubah simbol menjadi bentuk geometris yang dapat diamati [13]. Terdapat empat karakteristik utama dalam visualisasi menurut McCormick, yaitu penggunaan pola, perbandingan gambar, animasi, dan warna. Penggunaan pola membantu dalam proses pengenalan dan penarikan kesimpulan, sedangkan perbandingan gambar memungkinkan analisis berdasarkan perbedaan bentuk dan tekstur. Perbandingan animasi

berguna untuk memahami dinamika perubahan seiring waktu, dan perbandingan warna membantu membedakan informasi berdasarkan deskripsi visual.

2.4. Web Scraping

Web scraping adalah teknik pengambilan data secara otomatis dari situs web tanpa perlu melakukan penyalinan manual. Tujuan utama dari web scraping adalah untuk memperoleh informasi tertentu yang kemudian dapat dikumpulkan dan dianalisis secara terstruktur. Proses ini terdiri dari beberapa langkah penting, yaitu membuat template scraping untuk mengenali elemen HTML yang relevan, memahami navigasi situs web yang akan diekstrak, mengotomatisasi proses navigasi dan ekstraksi, serta menyimpan data yang diperoleh ke dalam database [14]. Dengan metode ini, data yang diambil menjadi lebih terfokus dan relevan, sehingga memudahkan dalam proses analisis atau pengembangan sistem berbasis data.

2.5. Question Answering

Sistem question answering (QA) merupakan salah satu pencapaian penting dalam bidang kecerdasan buatan karena memungkinkan pengguna untuk mengajukan pertanyaan dalam bahasa alami dan mendapatkan jawaban secara langsung [5]. Sistem QA terdiri dari beberapa komponen inti, seperti analisis pertanyaan (*question analysis*) yang bertugas memahami struktur sintaksis dan semantik pertanyaan, pembentukan kueri (*query generation*) untuk mencari informasi yang relevan, serta proses pembangkitan dan penilaian jawaban (*candidate generation and answer scoring*) yang menyeleksi dan memberi skor pada kandidat jawaban [15]. Berdasarkan jenis pertanyaan, QA dapat diklasifikasikan menjadi lima tipe utama: factoid, non-factoid, yes/no, list, dan opini. Namun, pada domain Bahasa Indonesia, sistem QA yang ada sebagian besar masih berfokus pada pertanyaan factoid seperti nama orang, tempat, atau tanggal [16].

2.6. Transformer

Transformer adalah arsitektur jaringan saraf yang menjadi terobosan dalam pemrosesan bahasa alami karena mampu mengatasi keterbatasan model sebelumnya seperti RNN dan CNN dalam memahami konteks jarak jauh [17]. Arsitektur ini menggunakan mekanisme *attention*, khususnya *self-attention*, untuk memberikan bobot pada kata-kata yang relevan dalam teks saat membentuk representasi. Salah satu komponen penting dalam Transformer adalah *multi-head attention* yang memungkinkan model menangkap informasi dari berbagai representasi konteks secara paralel. Selain itu, Transformer menggunakan struktur encoder dan decoder yang fleksibel, sehingga dapat memproses input dan output dengan panjang yang berbeda. Pendekatan ini meningkatkan efisiensi pemrosesan data sekuensial dan menjadi dasar pengembangan banyak model NLP modern seperti BERT dan GPT.

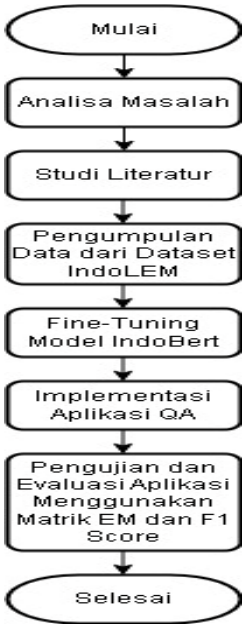
2.7. IndoBert

IndoBERT (Indo Bidirectional Encoder Representations From Transformer) merupakan modifikasi dari model BERT yang dirancang khusus untuk Bahasa Indonesia dengan arsitektur *BERT-base* yang memiliki 12 lapisan tersembunyi (*hidden layers*), 12 *attention heads*, dan 3.072 dimensi pada *feed-forward layers* [2][18]. Model ini dilatih menggunakan lebih dari 220 juta kata yang berasal dari Wikipedia, media berita seperti Kompas dan Tempo, serta korpus web Indonesia. Salah satu keunggulan IndoBERT adalah penerapan *max multi-head attention*, yaitu mekanisme yang mengambil nilai maksimum dari output berbagai *attention heads* untuk meningkatkan representasi teks [19]. Dibandingkan dengan model lain seperti CNN, RNN, LSTM, dan Bi-LSTM, IndoBERT memiliki kemampuan yang lebih baik dalam memahami konteks kata secara bidirectional, meskipun membutuhkan daya komputasi yang lebih besar. Penelitian terdahulu menunjukkan efektivitasnya, seperti pada analisis sentimen aplikasi mobile [7] dan aplikasi kesehatan [8]. Selain itu, benchmark seperti IndoNLU [18] dan dataset IndoLEM [2] digunakan untuk menguji performa model ini dalam berbagai tugas NLP Bahasa Indonesia.

3. METODE PENELITIAN

Metode dalam penelitian ini dimulai dengan analisis masalah untuk memahami kebutuhan sistem QA dan tantangan dalam menangani struktur Bahasa Indonesia. Selanjutnya, dilakukan studi literatur yang mencakup kajian terhadap model deep learning berbasis transformer, seperti BERT, serta implementasinya dalam domain QA. Setelah itu, data dikumpulkan dari dataset IndoLEM, yang dirancang khusus untuk evaluasi performa model NLP dalam Bahasa Indonesia. Proses inti melibatkan fine-tuning model IndoBERT menggunakan dataset IndoLEM untuk menyesuaikan model dengan tugas QA. Hasil fine-tuning ini diimplementasikan dalam sebuah aplikasi berbasis web, yang dirancang untuk memberikan jawaban atas pertanyaan berbasis teks secara akurat. Aplikasi ini. kemudian diuji menggunakan metrik evaluasi seperti Exact Match (EM) dan F1 Score untuk mengukur akurasi dan relevansi jawaban yang dihasilkan. Akhirnya, penelitian ini memberikan kesimpulan terkait efektivitas model yang diusulkan dan kontribusinya dalam pengembangan sistem QA berbasis Bahasa Indonesia. Arsitektur skema umum penelitian ini dapat dilihat pada gambar dibawah.

Penelitian ini bertujuan untuk mengembangkan sistem *Question Answering* (QA) berbasis Bahasa Indonesia dengan menerapkan pendekatan *transfer learning* menggunakan model IndoBERT dan dataset IndoLEM. Metode penelitian dilakukan melalui beberapa tahapan sistematis sebagaimana dijelaskan berikut.



Gambar 1. Diagram Alur Penelitian

3.1. Desain Penelitian

Penelitian ini menggunakan pendekatan eksperimen kuantitatif yang berfokus pada pengembangan dan evaluasi model QA berbasis *transformer*. Studi literatur dilakukan terlebih dahulu untuk mengidentifikasi teori dan penelitian terdahulu terkait NLP, QA, dan *transfer learning*. Hasil studi literatur menjadi dasar dalam perancangan arsitektur sistem dan proses pelatihan model.

3.2. Pengumpulan Data

Data utama yang digunakan dalam penelitian ini adalah dataset IndoLEM, yaitu kumpulan data terstruktur yang dirancang khusus untuk mengevaluasi model NLP Bahasa Indonesia. Dataset ini mencakup pasangan konteks dan pertanyaan yang digunakan dalam proses pelatihan (*training*), pengujian (*testing*), dan evaluasi sistem.

3.3. ra-pemrosesan Data

Data teks yang diperoleh dari dataset IndoLEM melalui beberapa tahap preprocessing untuk meningkatkan kualitas masukan model, meliputi:

a. *Case folding* (penyeragaman huruf kecil),
b. *Tokenizing* (pemisahan kata menjadi token),
c. *Stopwords removal* (penghapusan kata tidak bermakna),
d. *Stemming* (mengembalikan kata ke bentuk dasar).

Langkah ini bertujuan untuk menghilangkan *noise* dan membuat data lebih terstruktur sebelum digunakan dalam pelatihan model.

3.4. Perancangan dan Pengembangan Sistem

Sistem QA dirancang menggunakan arsitektur *Transformer* dengan model IndoBERT sebagai inti. Proses pengembangan melibatkan tahap fine-tuning model IndoBERT menggunakan data pelatihan dari IndoLEM untuk menyesuaikan model dengan tugas QA dalam Bahasa Indonesia. Selanjutnya, model diintegrasikan ke dalam aplikasi berbasis web menggunakan bahasa pemrograman Python dan framework Flask, sehingga pengguna dapat mengajukan pertanyaan berbasis teks dan memperoleh jawaban secara otomatis.

3.5. Evaluasi Model

Kinerja sistem dievaluasi menggunakan dua metrik utama:

a. Exact Match (EM): mengukur persentase jawaban model yang identik dengan jawaban referensi.
b. F1-Score: mengukur tingkat kesesuaian jawaban berdasarkan presisi dan *recall*.

Selain itu, sistem juga menampilkan nilai start logit dan end logit sebagai representasi keyakinan model terhadap rentang teks yang dianggap sebagai jawaban.

3.6. Implementasi dan Pengujian

Setelah pelatihan selesai, sistem diuji menggunakan skenario kasus nyata dengan konteks dan pertanyaan yang relevan. Hasil jawaban model dibandingkan dengan jawaban referensi untuk menilai tingkat akurasi dan relevansi sistem. Evaluasi dilakukan dengan menghitung nilai EM dan F1-Score serta menganalisis kesesuaian makna jawaban yang dihasilkan.

4. HASIL DAN PEMBAHASAN

4.1. Implementasi Sistem

Sistem *Question Answering* (QA) berbasis IndoBERT berhasil dikembangkan sesuai dengan rancangan penelitian. Implementasi dilakukan menggunakan bahasa pemrograman Python dengan framework Flask**,** dan diintegrasikan ke dalam aplikasi berbasis web yang memungkinkan pengguna memasukkan konteks teks dan pertanyaan secara langsung untuk memperoleh jawaban otomatis. Antarmuka sistem dirancang sederhana dan intuitif, terdiri dari kolom *context* untuk memasukkan teks sumber, kolom *question* untuk pertanyaan, serta tombol pencarian jawaban. Sistem juga dapat menampilkan beberapa alternatif jawaban beserta skor keyakinan model terhadap setiap jawaban.

4.2. Proses Kerja Sistem

Alur kerja sistem dimulai dari input teks konteks dan pertanyaan oleh pengguna. Data tersebut melalui tahap *preprocessing* (pembersihan teks, tokenisasi, dan penghapusan *stopwords*), kemudian diproses oleh model IndoBERT-QA untuk memahami hubungan semantik antara konteks dan pertanyaan. Model menentukan posisi awal (*start logit*) dan akhir (*end logit*) teks yang paling mungkin mengandung jawaban. Jawaban dengan nilai skor tertinggi akan ditampilkan sebagai hasil akhir, sedangkan beberapa jawaban

alternatif juga dapat diberikan untuk menunjukkan kemungkinan jawaban lain.

4.3. Pengujian dan Perhitungan Evaluasi

Evaluasi sistem dilakukan menggunakan metrik Exact Match (EM) dan F1-Score untuk mengukur kesesuaian hasil prediksi model terhadap jawaban referensi (*ground truth*). Selain itu, nilai start logit dan end logit juga dihitung untuk menunjukkan tingkat keyakinan model terhadap lokasi jawaban dalam konteks.

a. Contoh Perhitungan EM dan F1-Score
   - Jawaban referensi: "Phising adalah upaya mendapatkan data pribadi dengan menyamar sebagai pihak terpercaya."
   - Jawaban model: "Upaya memperoleh data pribadi."

b. Exact Match (EM):
   EM=1 jika jawaban 100% identik,
   EM=0 jika jawaban tidak 100% identik.
   Maka hasil EM dari jawaban **refensi** dan jawaban model diatas adalah 0, karena jawaban referensi dan jawaban model 100% tidak sama persis.

c. F1-Score:

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

   - *Precision(model)* = $\frac{jumlah\ overlap}{jumlah\ kata} = \frac{3}{4} = 0.75$
   - *Recall(referensi)*= $\frac{jumlah\ overlap}{jumlah\ kata} = \frac{3}{11} = 0.273$

   Maka;

$$F1 = \frac{2 * (0.75 * 0.273)}{0.75 + 0.273} = \frac{0.4095}{1.023} = 0.4$$

d. Perhitungan Skor Logit
   Model juga menghitung skor keyakinan dengan menjumlahkan nilai *start logit* dan *end logit* dari prediksi:
   Score=start_logit+end_logitScore = start_logit + end_logitScore=start_logit+end_logit
   Contoh:
   - *start logit* = 5.193
   - *end logit* = 5.570

$$Score = start\_logit + end\_logit$$

   Maka jawaban yang diberikan: Upaya memperoleh data pribadi
   score = 5.193 +5.570 = 10.763
   Semakin tinggi skor total logit, semakin tinggi pula keyakinan model terhadap jawaban yang diberikan.

Tabel 1. Kasus Uji coba

No	Pertanyaan	Jawaban Model	Jawaban Referensi	EM	F1
1	Apa itu phishing?	Upaya memperoleh data pribadi	Phising adalah upaya mendapatkan data pribadi dengan menyamar sebagai pihak terpercaya	0	0.40
2	Jelaskan tujuan serangan DDoS!	Membuat layanan tidak dapat diakses pengguna sah	Membuat layanan tidak dapat diakses pengguna dengan membanjiri server secara masif	0	0.87
3	Bagaimana modus penipuan online?	Memanfaatkan akun palsu dan penawaran murah	Dengan menggunakan akun palsu atau penawaran palsu lalu meminta transfer dana	0	0.50
4	Apa yang dimaksud dengan malware?	Perangkat lunak berbahaya	Perangkat lunak berbahaya yang dapat merusak atau mencuri data komputer korban	0	0.43

Hasil pengujian menunjukkan bahwa meskipun jawaban model sering kali tidak identik secara tekstual (EM = 0), nilai F1-Score tetap tinggi karena makna jawaban relevan dan sesuai dengan konteks. Hal ini menunjukkan bahwa model mampu memahami isi konteks dan mengekstrak informasi penting meskipun struktur kalimat berbeda.

4.4. Analisis Performa Sistem

Secara umum, sistem QA berbasis IndoBERT menunjukkan performa yang baik dalam menjawab pertanyaan berbasis teks Bahasa Indonesia. Model mampu memahami konteks secara semantik dan memberikan jawaban relevan dengan nilai F1-Score berkisar antara 0.40 hingga 0.87. Nilai *start-end logit* yang konsisten menunjukkan keyakinan tinggi model dalam memilih rentang teks yang tepat sebagai jawaban. Rendahnya nilai EM disebabkan oleh perbedaan redaksi antara jawaban model dan jawaban referensi, namun tidak mempengaruhi relevansi semantik jawaban. Hasil ini membuktikan bahwa pendekatan *transfer learning* menggunakan IndoBERT efektif untuk tugas *Question Answering* berbahasa Indonesia. Sistem tidak hanya mampu menghasilkan jawaban relevan tetapi juga memiliki potensi tinggi untuk diintegrasikan dalam berbagai aplikasi, seperti chatbot, asisten digital, dan sistem pencarian informasi otomatis.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan sistem *Question Answering* (QA) berbahasa Indonesia berbasis *transfer learning* dengan model IndoBERT dan dataset IndoLEM. Sistem mampu memahami konteks dan menghasilkan jawaban relevan dengan nilai *F1-Score* 0.40–0.87, meskipun nilai *Exact Match* masih rendah akibat perbedaan redaksi. Hasil ini membuktikan efektivitas IndoBERT dalam tugas QA Bahasa Indonesia. Untuk pengembangan selanjutnya, peningkatan dapat dilakukan dengan memperluas dataset, menerapkan *fine-tuning* lanjutan, serta mengembangkan antarmuka yang lebih interaktif agar

sistem lebih adaptif dan bermanfaat dalam berbagai aplikasi digital.

DAFTAR PUSTAKA

[1] Chandra, Y. W., & Suyanto, S. (2019). Indonesian Chatbot of University Admission Using a Question Answering System Based on Sequence-to-Sequence Model. *Procedia Computer Science*, *157*, 367–374. https://doi.org/10.1016/j.procs.2019.08.179

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

[3] Topsakal, O., & Akinci, T. C. (2023). Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast. *International Conference on Applied Engineering and Natural Sciences*, *1*(1), 1050–1056.

[4] Apriliani, D., Handayani, S. F., Anugrahaeni, T. N., Miftahudin, A., Nurarifiah, L., & Saputra, I. T. (2023). APLIKASI QUESTION ANSWER SEBAGAI MEDIA PEMBELAJARAN INTERAKTIF UNTUK MATA PELAJARAN AKUNTANSI. *JMM (Jurnal Masyarakat Mandiri)*, *7*(2), 2003.

[5] Agrawal, A., Atri, A., Chowdhury, A., Koneru, R., Batchu, K., & Mallavaram, S. (2021). Question Answering System Using Natural Language Processing. *International Journal of Research in Engineering, Science and Management*, *4*(12).

[6] Arifiyanti, A. A., Kartika, D. S. Y., & Prawiro, C. J. (2022). Using Pre-Trained Models for Sentiment Analysis in Indonesian Tweets. *2022 6th International Conference on Informatics and Computational Sciences (ICICoS)*, 78–83.

[7] Nugroho, K. S., Sukmadewa, A. Y., Wuswilahaken DW, H., Bachtiar, F. A., & Yudistira, N. (2021). BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews. *6th International Conference on Sustainable Information Engineering and Technology 2021*, 258–264.

[8] Imaduddin, H., A'la, F. Y., & Nugroho, Y. S. (2023). Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach. *International Journal of Advanced Computer Science and Applications*, *14*(8).

[9] Dhyani, B. (2021). Transfer Learning in Natural Language Processing: A Survey. *Mathematical Statistician and Engineering Applications*, *70*(1), 303–311.

[10] Muliyono, M. (2021). *Identifikasi Chatbot dalam Meningkatkan Pelayanan Online Menggunakan Metode Natural Language Processing* (Doctoral dissertation, Universitas Putra Indonesia YPTK).

[11] Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text Preprocessing For Student Complaint Document Classification Using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, *874*(1), 012017.

[12] Arabnia, H. (1999). Reading in information visualization: using vision to Think [Media Review]. *IEEE Multimedia*, *6*(4), 93–93.

[13] Marefat, M. M., Varecka, A. F., & Yost, J. (1997). An Intelligent Visualization Agent for simulation-based decision support. *IEEE Computational Science and Engineering*, *4*(3), 72–82.

[14] A. Yani, D. D., Pratiwi, H. S., & Muhardi, H. (2019). Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*, *7*(4), 257.

[15] Verberne, S., van Halteren, H., Theijssen, D., Raaijmakers, S., & Boves, L. (2011). Learning to rank for why-question answering. *Information Retrieval*, *14*(2), 107–132.

[16] Juliane, C., Armant, A. A., Sastramihardja, H. S., & Supriana, I. (2018). Question-answer pair templates based on bloom's revised taxonomy. *IOP Conference Series: Materials Science and Engineering*, *434*, 012281.

[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

[18] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5797–5808.

[19] A. S. Ekakristi, E. R. Nugroho, and S. B. Budi, "Intermediate-task transfer learning for Indonesian NLP tasks," *Computer Speech & Language*, vol. 83, 2025.

[20] M. Laugiwa and E. Yulianti, "Question Answering through Transfer Learning on Closed-Domain Educational Websites," *J. RESTI (Rekayasa Sistem & Teknologi Informasi)*, vol. 9, no. 1, pp. 104–110, Feb. 2025.

[21] N. Rachmawati and E. Yulianti, "Transfer Learning for Closed Domain Question Answering in COVID-19," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 13, no. 12, pp. 277–286, 2025.

[22] M. Namazifar, A. Papangelis, G. Tur, and D. Hakkani-Tür, "Language Model is All You Need: Natural Language Understanding as Question Answering," *arXiv preprint arXiv:2011.03023*, Nov. 2020.

[23] T. I. Ramadhan, A. Supriatman, dan T. R. Kurniawan, "Evaluasi dan Implementasi

IndoBERT Question Answering (QA) pada Domain Spesifik Menggunakan Mean Reciprocal Rank," *Jurnal Algoritma*, vol. 21, no. 1, hlm. 1–14, 2024.

[24] H. A. Pandya, B. Ardeshna, dan B. S. Bhatt, "Cascading Adaptors to Leverage English Data to Improve Performance of Question Answering for Low-Resource Languages," *arXiv preprint arXiv:2112.09866*, 2021.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, dan P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020.

[26] T. I. Ramadhan, A. Supriatman, dan T. R. Kurniawan, "Evaluasi dan Implementasi IndoBERT Question Answering (QA) pada Domain Spesifik Menggunakan Mean Reciprocal Rank," *Jurnal Algoritma*, vol. 21, no. 1, hlm. 1–14, 2024.

[27] D. Suhartono, "Towards automatic question generation using pre-trained models for Bahasa Indonesia," *Educ. Inf. Technol.* / Springer (article), 2024.

[28] A. Hodijah, A. Izzati, dan S. A. Susanti, "Analysis Ontology-Based Simple Closed Domains Question," *Atlantis Press (conf. proc.)*, 2024.

[29] Siregar, F. A., Siregar, A. F., & Setiadi, E. W. N. (2024). Sistem Pendukung Keputusan Menentukan E-Commerce Terbaik Menggunakan Metode Topsis. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, *5*(3), 935-947. https://pkm.tunasbangsa.ac.id/index.php/kesatria/article/view/419.