LOGIC: Jurnal Penelitian Informatika
Vol. 2, Issue 1, pp. 12–17 (2024)
doi: http:/doi.org/10.25124/logic.v2i1.7437

RESEARCH ARTICLE

Identifikasi Similar Question dengan IndoBERT (Studi Kasus Dataset QAS Covid-19)

Rifki Adi Pramana and Ade Romadhony*

Fakultas Informatika, Universitas Telkom, Bandung, 40257, Jawa Barat, Indonesia
*Corresponding author: aderomadhony@telkomuniversity.ac.id
Received on 13 April 2024; accepted on 19 May 2024

Abstrak

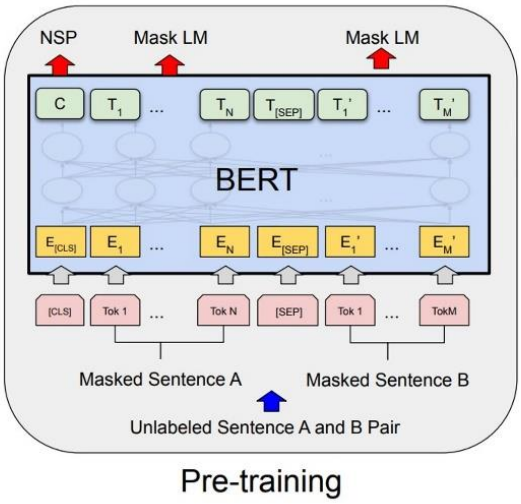
Question answering system (QAS) merupakan sebuah *task* pada bidang informatika, secara lebih spesifik yaitu pada bidang *Natural Language Processing* (NLP). Sebuah QAS menyediakan jawaban secara otomatis berdasarkan pertanyaan yang diberikan oleh pengguna. Salah satu bagian dari tahapan pemrosesan dalam QAS adalah identifikasi pertanyaan yang mirip (*similar question identifition*). Tahapan similar question identification bertujuan untuk mengidentifikasi pertanyaan yang mirip, sehingga didapatkan jawaban yang tepat. Pada penelitian ini, dilakukan identifikasi *similar question* pada dataset yang berisi pertanyaan seputar Covid-19. Identifikasi similar question diaplikasikan dengan memanfaatkan model IndoBERT, dimana diterapkan pengukuran *similarity* berdasarkan *cosine similarity*. Berdasarkan eksperimen yang dilakukan, diperoleh 197 dari total 611 pasang pertanyaan yang berhasil diidentifikasi kemiripannya. Analisis terhadap hasil identifikasi menunjukkan bahwa faktor yang memperngaruhi dalam kemiripan antar pertanyaan antara lain adalah panjang dari suatu kalimat yang dibandingkan, kata awal dari kalimat yang dibandingkan, dan relevansi antar beberapa kata yang terdeteksi memiliki kemiripan satu sama lain.

Key words: pemrosesan bahasa alami, *question similarity*, IndoBERT, Covid-19

Pendahuluan

Coronavirus merupakan bagian dari keluarga besar virus yang menyebabkan infeksi saluran pernafasan. Coronavirus ini baru ditemukan pada Desember 2019, sehingga penyakit ini disebut Coronavirus Disease-2019 (Covid-19)[1]. Karena pandemi ini tergolong baru maka rasa ingin tahu untuk menelusuri lebih jauh terkait Covid19 pun muncul di kalangan masyarakat. Dengan semakin berkembangnya teknologi, maka pencarian terkait Covid-19 akan semakin mudah, salah satunya dengan menggunakan sistem question answering. Question Answering Systems (QAS) adalah salah satu bidang pada ilmu komputer yang mengolah dokumen teks dan pengambilan informasi sebagai aspek penting. Ide utama pada QAS yaitu berbasis pengetahuan yang terdiri dari pencarian informasi yang diminta oleh pengguna yang mengekspresikan diri menggunakan Natural Language Processing (NLP)[2]. Secara garis besar, QAS dapat juga diartikan dengan sistem pencarian informasi yang mengharapkan pertanyaan yang diajukan untuk dijawab dengan benar atau dijawab secara langsung[3]. Perkembangan pada NLP mempengaruhi QAS, dimana dulu hanya dapat menjawab pertanyaan secara terbatas dalam satu bidang berdasarkan informasi yang terstruktur, namun sekarang sistem ini sudah dapat menjawab berbagai pertanyaan dengan sumber informasi yang tidak terstruktur[4]. Untuk mengidentifikasi pertanyaan, digunakan mekanisme yang disebut Question Similarity mechanism. Mekansime ini akan menghitung cosine similarity antara pertanyaan-pertanyaan yang diajukan. QAS menggunakan mekanisme Question Similarity sebagai filter pertanyaan[5]. Mekanisme ini dapat dilakukan dengan menggunakan berbagai macam model, seperti word embedding models berupa Word2Vec ataupun Global 2 Vectors for Word Representation (GloVe), Bidirectional Encoder Representations from Transformers (BERT), dsb. IndoBERT merupakan sebuah model berbasis transformers yang berupa BERT, namun dilatih untuk dijadikan sebagai model bahasa berjenis masked language model menggunakan kerangka kerja Huggingface[6], yang secara spesifik dkhususkan untuk megidentifikasi kata ataupun kalimat dalam bahasa Indonesia. BERT sendiri adalah model Bahasa yang telah dilatih sebelumnya pada sejumlah besar teks tanpa label yang mencapai performa terbaik dalam berbagai tugas NLP[7]. Secara efektif, BERT mengodekan inputan teks untuk dilakukan pre-trained menggunakan model bahasa pada sebuah korpus mentah yang besar. dan kemudian disesuaikan (finetuned) untuk setiap tugas spesifik, termasuk klasifikasi kalimat, klasifikasi pasangan kalimat, dan menjawab pertanyaan. Dikarenakan telah dilakukan pre-trained pada korpus yang besar, BERT dapat mencapai akurasi yang tinggi bahkan jika ukuran data

untuk tugas tertentu tidak cukup besar[8]. Pada penelitian ini, dilakukan identifikasi similarity score setiap pasang pertanyaan yang mewakili sebuah topik dan klaster pertanyaan, untuk kemudian dicari pasangan pertanyaan dengan skor tertinggi dan melakukan perbandingan dengan pasangan topik pertanyaan dan klaster pertanyaan dari dataset. Jika pasangan pertanyaan hasil identifikasi yang memiliki skor tertinggi sama dengan pasangan pertanyaan pada dataset, maka akan diberi label 1, sebaliknya jika pasangan pertanyaan hasil identifikasi yang memiliki skor tertinggi tidak sama dengan pasangan pertanyaan pada dataset akan diberi label 0. Keakuratan model akan dihitung berdasarkan banyaknya label 1, yaitu pasangan pertanyaan hasil identifikasi yang sama dengan pasangan pertanyaan pada dataset. Peneltian ini memanfaatkan model IndoBERT sebagai representasi pertanyaan, sebagai dasar pengukuran nilai similarity. Model ini digunakan karena IndoBERT merupakan model monolingual BERT yang dapat digunakan khusus untuk mengidentifikasi kalimat – kalimat dalam Bahasa Indonesia, dan memiliki performa keakuratan yang baik dibandinkan model multilingual BERT atau model lainnya.



Gambar 1. Arsitektur BERT [11]

Tinjauan Pustaka

Studi Terkait

Menghitung bagaimana kata – kata muncul secara bersamaan berdasarkan rangkaian kata yang ada, merupakan cara untuk mengukur question similarity secara langsung. Perbedaan utama terletak pada metode kalkulasi yang berbeda. Terdapat metode untuk menemukan urutan umum yang terpanjang, jarak pengeditan minimum, dan n-gram pada rangkaian kata atau setiap korpusnya. Metode-metode ini memiliki prinsip yang sederhana, mudah untuk diimplementasikan, dan dapat langsung membandingkan teks asli. Dalam hal lain, mengukur jarak antar teks dengan mengkalkulasikan ruang vektor dapat dilakukan juga oleh question similarity berdasarkan metode statistik. Semakin besar jaraknya, maka akan semakin rendah korelasi antar keduanya. Sebaliknya, semakin kecil jaraknya, maka akan semakin tinggi tingkat kemiripan antar keduanya. Saat ini, metode yang berdasarkan pada rangkaian kata jarang digunakan untuk menghitung question similarity. Namun, hasil dari metode-metode ini diintegrasikan ke dalam metode yang lebih kompleks. Deep learning telah menjadi metode yang paling popular dalam NLP, khususnya di bidang question similarity [9].

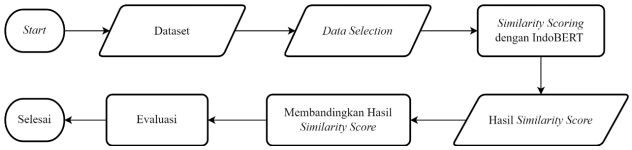
IndoBERT merupakan hasil pengembangan dari model BERT yang ditujukan untuk Bahasa Indonesia. Model ini dihasilkan melalui pelatihan dengan menggunakan dataset Indo4B. Indo4B adalah koleksi data berukuran besar yang terdiri dari sekitar empat miliar kata yang diambil dari berbagai sumber teks berbahasa Indonesia yang telah diolah terlebih dahulu. Dataset ini mencakup teks dari beragam sumber seperti berita daring, media sosial, Wikipedia, artikel daring, transkrip video, dan dataset secara paralel. Data tersebut mencakup kalimat-kalimat Bahasa Indonesia, baik yang formal maupun santai. Kedua model, baik IndoBERT maupun BERT, bergantung pada arsitektur transformer. Transformer adalah jenis model transduksi urutan yang inovatif, menggantikan pendekatan berulang yang biasanya digunakan dalam struktur encoder-decoder dengan mekanisme multi-headed self-attention. Arsitektur transformer terdiri dari masing-masing enam lapisan encoder dan decoder. Model transformer ini sangat efektif dalam menangani masalah penerjemahan pada mesin. Baik encoder maupun decoder, keduanya berupaya untuk memahami bahasa, dan hal itulah yang membuat model transformer menjadi sangat istimewa [10]. Arsitektur dari model BERT terlampir pada Gambar **??**.

IndoBERT menerapkan pendekatan pelatihan Masked Language Model (MLM) dan Next Sentence Prediction (NSP), mirip dengan cara BERT asli beroperasi. Proses pre-trained pada BERT dilakukan dengan mengajari model untuk menyelesaikan tugas MLM dan NSP. Tahap pre-trained yang melibatkan kedua tugas ini membantu BERT memahami konteks dan arti dalam bahasa. Pada tugas MLM, kata-kata yang diberi masked di sekitarnya digunakan untuk meramalkan kata yang seharusnya mengisi ruang kosong dalam kalimat tersebut. Karena BERT mampu memahami kata-kata secara bidirectional (dari arah kiri maupun kanan), hal ini membantu model dalam menebak kata-kata yang diberi masked. Sedangkan NSP bertujuan melatih model agar memahami hubungan antara dua kalimat yang ada [10].

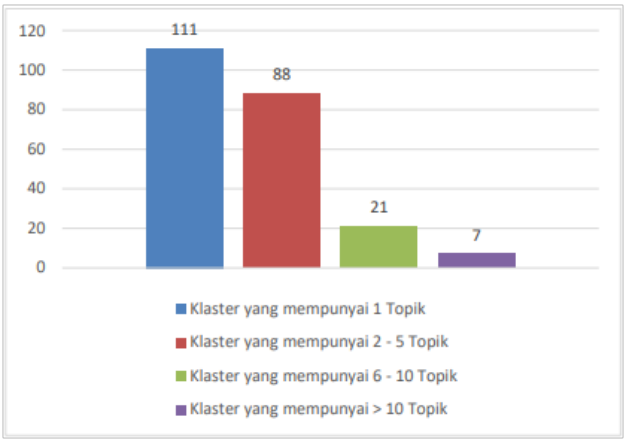
Pada penelitian yang dilakukan oleh M. Z. Aonillah dan koleganya di tahun 2022, dengan menggunakan model Recognizing Question Entailment (RQE) yang terdiri dari beberapa algoritma supervised learning, membangun Question Answering System (QAS) terkait pertanyaan – pertanyaan mengenai Covid-19 di Indonesia. Penelitian ini mencari similar questions dari pertanyaan pada dataset terhadap pertanyaan baru. Hasil penelitian dengan menggunakan algoritma Logistic Regression berdasarkan data uji yang terdiri 218 dari 725 total keseluruhan data, didapati sebanyak 184 data yang sesuai dan 34 data yang tidak sesuai, sehingga dapat diketahui keakuratan pada data ujinya sebanyak 84% [12]. Tujuan menggunakan penelitian ini sebagai referensi, karena terdapat perhitungan similar questions dengan penggunaan dataset yang hampir serupa dalam pembangunan QASnya secara keseluruhan.

Masih berkaitan dengan question similarity, pada tahun 2019 F. Kunneman dan koleganya mendeteksi question similarity pada forum Community Question Answering (CQA) dengan menggunakan berbagai macam model. Model yang digunakan yaitu BM25, Translation-Based Language Model (TRLM), SoftCosine, Smoothed Partial Tree Kernels (SPTK), dan Ensemble. Berdasarkan hasil yang didapat, model yang memiliki performansi terbaik adalah SoftCosine dan Ensemble yang mengimplementasikan kombinasi metriks Word2Vec dan ELMo, dengan nilai Mean Average Precision (MAP) masing – masing modelnya 73,89 dan 73,90. Berdasarkan penelitian ini juga, dapat diketahui bahwa pemilihan tahap preprocessing dan penggunaan word distributions dapat mempengaruhi performansi model [13].

Salah satu model yang dapat dimanfaatkan untuk mengidentifikasi pertanyaan – pertanyaan ataupun kalimat yang spesifik menggunakan Bahasa Indonesia adalah IndoBERT. Penelitian terkait model ini sudah banyak dilakukan, seperti yang dilakukan oleh S. M. Isa dan koleganya pada tahun 2021. Penelitian tersebut dilakukan untuk mendeteksi berita palsu yang ada di Indonesia dengan menggunakan IndoBERT. Model



Gambar 2. Flowchart sistem yang dibangun



Gambar 3. Distribusi Data Berdasarkan Klaster

ini menghasilkan akurasi sebesar 94,66% pada dataset yang terdiri dari 3.465 berita palsu dan 766 berita asli. Pada penelitian ini juga terdapat perbandingan model dengan TF-IDF + SVM dan TF-IDF + Naïve Bayes. Hasil Akurasi yang dihasilkan dari kedua model ini masing – m

Metodologi Penelitian

Pada penelitian ini, terdapat beberapa tahapan yang dilakukan oleh penulis. Alur dari tahapan yang dilakukan oleh penulis dalam penelitian ini terlampir pada Gambar 2.

Dataset

Pada penelitian ini, dataset yang digunakan adalah dataset dari penelitian sebelumnya yang dilakukan oleh M. Z. Aonillah di tahun 2022 [12], berupa kumpulan pertanyaan dan jawaban seputar Covid-19 yang diperoleh dari akun Twitter (sekarang berubah namanya menjadi media sosial X) yang berkredibilitas seperti dokter, satgas Covid-19, dan Kemenkes RI. Jawaban dari pertanyaan yang ada pada dataset diambil dari berbagai macam website resmi di bidang kesehatan seperti website milik WHO, Alodokter, dan sebagainya. Dataset ini memiliki total 725 pasang data yang terdiri dari teks pertanyaan, topik pertanyaan, klaster pertanyaan, kategori pertanyaan, dan jawaban pertanyaan, yang sudah didefinisikan pada Tabel 1 di bab sebelumnya. Adapun untuk distribusi data berdasarkan klasternya dapat dilihat pada Gambar 3.

Dimana terdapat sebanyak 111 klaster pertanyaan yang hanya memiliki 1 topik pertanyaan, terdapat 88 klaster pertanyaan yang memiliki 2 sampai dengan 5 topik pertanyaan, terdapat 21 klaster pertanyaan yang memiliki 6 sampai dengan 10 topik pertanyaan, dan terdapat 7 klaster pertanyaan yang memiliki topik pertanyaan lebih dari 10.

Data Selection

Pada tahap ini, penulis melakukan seleksi terhadap dataset yang ada. Dari jenis label atau kolom pada dataset yang terlampir pada Tabel 1, yang digunakan pada penelitian ini adalah kolom Topik Pertanyaan dan Klaster Pertanyaan. Topik Pertanyaan yang dipakai ada sebanyak 611 pertanyaan, dan Klaster Pertanyaan yang dipakai ada sebanyak 227 pertanyaan. Topik Pertanyaan dan Klaster Pertanyaan merupakan pasangan pertanyaan yang saling berhubungan seperti yang terlampir pada Tabel 1.

Table 1. Contoh Topik Pertanyaan dan Klaster Pertanyaan

Topik Pertanyaan	Klaster Pertanyaan
Jika sudah terkena covid apakah masih harus vaksin ?	Apakah perlu vaksin bagi pasien ex COVID-19?
Apakah hidung mampet badan pegel-pegel dan deman 38° merupakan gejala covid-19?	Apa gejala dari COVID-19?
Apakah menggigil dan gampang kedinginan merupakan gejala covid-19?	Apa gejala dari COVID-19?
Apakah melakukan olahraga dapat terpapar COVID-19?	Aktifitas apa saja yang ber-resiko terkena COVID-19?

Similarity Scoring Berdasarkan Representasi IndoBERT

Proses pemodelan IndoBERT untuk word embedding melibatkan serangkaian langkah yang mendalam dalam mengubah teks bahasa Indonesia menjadi representasi numerik yang kaya dengan informasi semantik. Pertama, teks preprocessing dilakukan, termasuk langkah-langkah seperti data cleansing dan tokenization, di mana teks dipecah menjadi unit-unit lebih kecil yang disebut token. Setelah tokenization, model IndoBERT dimuat menggunakan library Hugging Face Transformers. Setiap token kemudian diteruskan melalui model IndoBERT. Model ini memiliki lapisan-lapisan transformer yang berfungsi untuk memproses teks secara kontekstual. Pada setiap lapisan, token-token diolah berdasarkan konteks kata-kata sebelumnya dan sesudahnya. Representasi vektor yang dihasilkan oleh model ini merefleksikan pemahaman tentang hubungan semantik dan konteks antara kata-kata dalam teks. Selanjutnya, vektor-vektor representasi ini diambil sebagai word embeddings. Artinya, setiap kata dalam teks direpresentasikan oleh vektor numerik multidimensi. Representasi ini mencakup makna kata itu sendiri dan cara kata tersebut berinteraksi dengan kata-kata lain dalam konteks teks. Word embeddings ini dapat digunakan untuk berbagai tugas NLP, seperti analisis sentimen, klasifikasi teks, pemahaman entitas bernama, dan tugas lainnya. Selain itu, pengguna juga dapat mengekstraksi fitur dari lapisan-lapisan tertentu dalam model untuk tujuan analisis yang lebih dalam [11, 15, 16].

Pada penelitian ini, proses pemodelan IndoBERT untuk word embedding yang akan menghasilkan representasi vektor melibatkan tahapan preprocessing berupa punctual removal dan multiple spaces removal, tokenization, pemrosesan oleh model IndoBERT dengan 2 pretraining phase, yaitu Indobertbase-p1 dan Indobert-base-p2. Pretraining phase pertama dilatih dengan total waktu 35 jam, dengan panjang sequence maksimum sebanyak 128. Sedangkan pretraining phase kedua dilatih dengan total waktu 9 jam dengan panjang sequence maksimum sebanyak 512 [17]. Pemanfaatan representasi vektor ini akan digunakan untuk memprediksi perhitungan similarity score dari setiap inputan pertanyaan. Perhitungan tersebut dilakukan dengan menggunakan cosine similarity yang mengukur nilai antara dua vektor dalam ruang multidimensi, sesuai dengan persamaan **??**.

$$\text{CosineSimilarity}\ (x, y) = \frac{\sum_{i=1}^{n} x[i] \cdot y[i]}{\|x\| \cdot \|y\|} = \frac{\sum_{i=1}^{n} x[i] \cdot y[i]}{\sqrt{\sum_{i=1}^{n} (x[i])^2} \cdot \sqrt{\sum_{i=1}^{n} (y[i])^2}}$$

Table 2. Ilustrasi Sistem Prediksi Similarity Score

Topik Pertanyaan	Klaster Pertanyaan
Apakah rapid test dapat membaca virus lain?	Apa gejala dari COVID-19?
Apakah rapid test dapat membaca virus lain?	Apakah test COVID-19 dapat mendeteksi virus lain?
Apakah rapid test dapat membaca virus lain?	Apa saja jenis vaksin COVID-19?
apa saja obat untuk covid-19	Apa gejala dari COVID-19?
apa saja obat untuk covid-19	Apakah test COVID-19 dapat mendeteksi virus lain?
apa saja obat untuk covid-19	Apa saja jenis vaksin COVID-19?

Table 3. Ilustrasi Identifikasi Similarity Score Tertinggi

Topik Pertanyaan	Klaster Pertanyaan
Apakah rapid test dapat membaca virus lain?	Apakah test COVID-19 dapat mendeteksi virus lain?
apa saja obat untuk covid-19	Apa saja jenis vaksin COVID-19?

Dimana x_i merupakan representasi vektor dari kalimat masukan pertama, dan y_i merupakan representasi vektor dari kalimat masukan kedua yang ingin dibandingkan. Sementara, i menunjukkan indeks dari representasi vektor di tiap kalimat. Dan n merupakan banyaknya indeks dari representasi vektor panjangya kalimat [18].

Membandingkan Hasil Similarity Score Tertinggi dari Topik dan Klaster Pertanyaan

Berdasarkan proses pengolahan data yang telah dilakukan sebelumnya, didapatkan similarity score dengan pemodelan IndoBERT. Sistematis proses prediksi similarity score pada tahapan sebelumnya yaitu membandingkan setiap pertanyaan dari 'Topik Pertanyaan' dengan seluruh pertanyaan pada 'Klaster Pertanyaan'. Analogi dari sistem tersebut terlampir pada Tabel 2.

Dari tiap perbandingan Topik dan Klaster Pertanyaan didapatkan hasil similarity score masing-masing. Kemudian dari tiap 'Topik Pertanyaan' diambil hasil similarity score terhadap 'Klaster Pertanyaan' yang paling tinggi. Contoh hasil similarity score tertinggi dari tiap 'Topik Pertanyaan' terlampir pada Tabel 3.

Evaluasi

Evaluasi dilakukan dengan menyesuaikan 'Klaster Pertanyaan' dari hasil identifikasi dengan 'Klaster Pertanyaan' dari dataset pada setiap 'Topik Pertanyaan' yang ada. Kemudian dilakukan pelabelan terhadap 'Klaster Pertanyaan' hasil identifikasi yang sesuai dengan 'Klaster Pertanyaan' pada dataset di setiap topiknya. Jika sesuai maka akan diberi label 1. Sebaliknya, jika tidak sesuai maka akan diberi label 0.

Persentase keakuratan hasil identifikasi dapat dilihat dengan menghitung banyaknya label 1 yang dihasilkan, dengan menggunakan persamaan **??**. Dimana, acc merupakan akurasi yang ingin dihitung, x merupakan jumlah hasil identifikasi yang benar atau berlabel 1, dan y merupakan total banyaknya dataset yang ada.

$$acc = \frac{x}{y} \cdot 100\% \quad (1)$$

Hasil dan Pembahasan

Proses Identifikasi Similarity Score pada Seluruh Dataset

Setelah melalui proses identifikasi dengan menggunakan model IndoBERT, maka didapatkan similarity score dari masing-masing 'Topik Pertanyaan' dengan seluruh 'Klaster Pertanyaan' yang ada pada dataset. Total data yang dihasilkan pada proses ini ada sebanyak 138.697 data, yang diperoleh dari 611 'Topik Pertanyaan' dan 227 'Klaster Pertanyaan'. Pada Tabel 4 ditampilkan beberapa hasil dari skema identifikasi similar question seperti yang sudah dijelaskan.

Setelah didapatkan similarity score dari skema pengecekan tersebut, selanjutnya akan dilakukan data filtering, yang mana hanya akan mengambil pasangan antara tiap 'Topik Pertanyaan' dan 'Klaster Pertanyaan' dengan similarity score tertinggi. Tabel 5 merupakan contoh data yang diambil dengan similarity score tertinggi pada tiap 'Topik Pertanyaan' yang ada.

Perbandingan antara Hasil Similarity Check dengan Label Dataset

Hasil data filtering seperti contoh pada Tabel 5, akan diidentifikasi lebih lanjut dengan cara dibandingkan dengan dataset awal. Pada identifikasi ini, akan dilihat dari tiap 'Topik Pertanyaan', apakah 'Klaster Pertanyaan' dari dataset awal memiliki kesamaan dengan 'Klaster Pertanyaan' dari hasil pengecekan similar question yang sudah dilakukan. Beberapa contoh hasil identifikasi ini ditampilkan pada Tabel 6.

Hasil Analisis

Untuk pertanyaan yang berhasil diidentifikasi serupa memiliki rentang skor similarity dari 0,61 sampai dengan 1, sedangkan untuk pertanyaan yang diidentifikasi tidak serupa memiliki rentang skor similarity dari 0,60 sampai dengan 0,98. Confusion matrix dari penelitian ini terlampir pada Tabel 7.

Hasil analisis berdasarkan confusion matrix pada penelitian ini, didapati 197 pertanyaan yang berhasil diidentifikasi serupa dari total 611 pertanyaan yang sudah melewati beberapa tahapan sebelumnya, sehingga didapati hasil keakuratan identifikasi dari pemanfaatan representasi model IndoBERT sebesar 32%. Adapun, sebanyak 414 data diidentifikasi tidak serupa. Beberapa contohnya terlampir pada Tabel 8.

Berdasarkan poin No. 1 pada Tabel 8, dapat dilihat kemungkinan terjadinya ketidaksesuaian hasil identifikasi, dapat disebabkan karena model mengidentifikasi similar question berdasarkan panjang dari suatu kalimat. Karena tahapan pada word embedding, akan mengubah kalimat menjadi vektor untuk mengecek similar question.

Berdasarkan poin No. 2 pada Tabel 8, dapat dilihat bahwa model memiliki kemungkinan besar dalam melakukan pengecekan similar question berdasarkan kata awal pada suatu kalimat. Selain itu, adanya relevansi antara kata "obat" pada topik pertanyaan dan kata "vaksin" pada klaster pertanyaan hasil similarity check, merupakan salah satu faktor yang membuat kedua pertanyaan tersebut terdeteksi memiliki kemiripan antara satu sama lain.

Berdasarkan poin No. 3 pada Tabel 8, dapat dilihat bahwa model memiliki kemungkinan besar dalam mendeteksi adanya relevansi antara kata "2 minggu" pada topik pertanyaan dan kata "berapa lama" pada klaster pertanyaan hasil similarity check. Dua ungkapan tersebut memiliki relevansi berupa pertanyaan terkait waktu pada gejala Covid-19.

Berdasarkan poin No. 4 pada Tabel 8, saat dianalisis kembali, ternyata ditemukan bahwa hasil identifikasi klaster yang diberikan oleh model lebih cocok jika dibandingkan dengan klaster yang ada pada dataset. Sehingga ketidaksesuaian hasil dapat disebabkan juga karena dataset yang kurang spesifik dalam memasangkan topik dan klaster pertanyaan.

Menurut M. Z. Aonillah [12], hal lain yang dapat menyebabkan perbedaan pada hasil pengecekan similar question yaitu dikarenakan

Table 4. Contoh Pengukuran Similarity antar Pasangan Topik dan Klaster Pertanyaan

No.	Topik Pertanyaan	Klaster Pertanyaan	Similarity Score
1	Apakah rapid test dapat membaca virus lain?	Apa gejala dari COVID-19?	0.56793153
	Apakah rapid test dapat membaca virus lain?	Apakah test COVID-19 dapat mendeteksi virus lain?	0.86117524
	Apakah rapid test dapat membaca virus lain?	Apa saja jenis vaksin COVID-19?	0.5429102
2	apa saja obat untuk covid-19	Apa gejala dari COVID-19?	0.78230494
	apa saja obat untuk covid-19	Apakah test COVID-19 dapat mendeteksi virus lain?	0.7006149
	apa saja obat untuk covid-19	Apa saja jenis vaksin COVID-19?	0.8560339
3	Apakah sesak nafas merupakan gejala covid-19?	Apa gejala dari COVID-19?	0.82488585
	Apakah sesak nafas merupakan gejala covid-19?	Apakah test COVID-19 dapat mendeteksi virus lain?	0.67859995
	Apakah sesak nafas merupakan gejala covid-19?	Apa saja jenis vaksin COVID-19?	0.61707413

Table 5. Contoh Similar Questions Tertinggi dari Setiap Topik Pertanyaan

Topik Pertanyaan	Klaster Pertanyaan	Similarity Score
Apakah rapid test dapat membaca virus lain?	Apakah test COVID-19 dapat mendeteksi virus lain?	0.86117524
apa saja obat untuk covid-19	Apa saja jenis vaksin COVID-19?	0.8560339
Apakah sesak nafas merupakan gejala covid-19?	Apa gejala dari COVID-19?	0.82488585

adanya kesalahan penulisan dalam suatu kata, ataupun adanya kata yang tidak sesuai dengan ejaan yang sebenarnya.

Kesimpulan

Hasil dari keseluruhan eksperimen dan analisis ini menunjukkan bahwa IndoBERT merupakan suatu model yang cukup detail dalam melakukan pengecekan *similarity*. Dengan dataset yang digunakan dalam penelitian ini, IndoBERT berhasil mengidentifikasi sebanyak 197 pasang similar questions yang sama dengan dataset awal dari total 611 pasang data yang ada pada dataset. Faktor utama yang menyebabkan kecilnya jumlah data hasil identifikasi yaitu dikarenakan model IndoBERT memiliki performa yang sangat signifikan dalam pengecekan tiap detail kata yang ada pada suatu kalimat. Khususnya pada pengecekan kata awal antar satu kalimat dengan kalimat lain, pengecekan relevansi antar kata yang ada pada satu kalimat dengan kalimat lain, dan pengecekan terhadap penulisan kata pada suatu kalimat. Selain itu, faktor lain yang menyebabkan kecilnya hasil identifikasi dikarenakan dataset yang kurang spesifik dalam memasangkan topik dan klaster pertanyaannya. Akan ada banyak kemungkinan yang terjadi dalam eksperimen lebih lanjut terkait analisis similar questions. Penelitian ini difokuskan pada penggunaan model IndoBERT untuk menganalisis similar questions terkait Covid-19. Untuk penelitian lebih lanjut, dapat dilakukan eksperimen dengan membandingan penggunaan model IndoBERT dengan model lainnya dalam mengidentifikasi *similar questions*, ataupun dapat menggunakan dataset lainnya. Model lain ataupun dataset lain dapat menjadi fokus untuk penelitian lebih lanjut terkait identifikasi *similar questions*

Daftar Pustaka

1. Nasution NH. GAMBARAN PENGETAHUAN MASYARAKAT TENTANG PENCEGAHAN COVID-19 DI KECAMATAN PADANGSIDIMPUANBATUNADUA, KOTA PADANGSIDIMPUAN. 2021.
2. da Silva JWF, Venceslau ADP, Sales JE, Maia JGR, Pinheiro VCM, Vidal VMP. A short survey on end-to-end simple question answering systems. Artif Intell Rev. 2020 Oct;53(7):5429-53.
3. Sina DI, Romadhony A. Eksplorasi Reading Comprehension Berbasis Open Information Extraction Bahasa Indonesia.
4. Dan G, Lovina G, Informatika JT, Tinggi S, Surabaya T. QUESTION ANSWERING SYSTEM DAN PENERAPANNYA PADA ALKITAB; unknown. http://www.petra.ac.id/~puslit/journals/dir.php?DepartmentID=INF.
5. Aithal SG, Rao AB, Singh S. Automatic question-answer pairs generation and question similarity mechanism in question answering system. Applied Intelligence. 2021 Nov;51(11):8484-97.
6. Koto F, Rahimi A, Lau JH, Baldwin T. IndoLEM and IndoBERT: A Benchmark Dataset and Pretrained Language Model for Indonesian NLP. 2020.
7. Alshammari W, Alhumoud S. TAQS: An Arabic Question Similarity System Using Transfer Learning of BERT with BiLSTM. IEEE Access. 2022.
8. Sakata W, Tanaka R, Shibata T, Kurohashi S. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. In: SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. Association for Computing Machinery, Inc; 2019. p. 1113-6.

Table 6. Hasil Pelabelan Perbandingan Klaster Pertanyaan

Topik Pertanyaan	Klaster Pertanyaan (Dataset)	Klaster Pertanyaan (Hasil Similarity Check)	Similarity Score	Status
Apakah rapid test dapat membaca virus lain?	Apakah test COVID-19 dapat mendeteksi virus lain?	Apakah test COVID-19 dapat mendeteksi virus lain?	0.861175239	Sama
apa saja obat untuk covid-19	Obat atau vitamin apa saja yang dapat digunakan untuk membantu pemulihan	Apa saja jenis vaksin COVID-19?	0.856033921	Tidak Sama
Apakah sesak nafas merupakan gejala covid-19?	Apa gejala dari COVID-19?	Apa gejala dari COVID-19?	0.824885845	Sama
bagaimana cara ibu hamil agar tidak tertular covid-19	Bagaimana cara penanganan COVID-19 bagi ibu hamil?	Bagaimana cara penanganan COVID-19 bagi ibu hamil?	0.853228986	Sama
apakah pemerintah telah melakukan penilitian mengenai obat covid?	Obat atau vitamin apa saja yang dapat digunakan untuk membantu pemulihan COVID-19?	Apakah vaksin dibutuhkan untuk menagani COVID-19?	0.689313889	Tidak Sama

Table 7. Confusion Matrix Hasil Penelitian

		Predicted	
		True	False
Actual	True	197	414

9. Yuan Y, Zhang G. High school math text similarity studies based on CNN and BiLSTM. In: Proceedings - 2020 5th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2020. Institute of Electrical and Electronics Engineers Inc.; 2020. p. 1982-6.
10. Isa SM, Nico G, Permana M. INDOBERT FOR INDONESIAN FAKE NEWS DETECTION. ICIC Express Letters. 2022 Mar;16(3):289-97.
11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.
12. Aonillah MZ, Hasmawati H, Romadhony A. Question Entailment on Developing Indonesian Covid-19 Question Answering System. Journal of Computer System and Informatics (JoSYC). 2022 Sep;3(4).
13. Kunneman F, Ferreira TC, Krahmer E, Bosch AVD. Question similarity in community question answering: A systematic exploration of preprocessing methods and models. In: International Conference Recent Advances in Natural Language Processing, RANLP. Incoma Ltd; 2019. p. 593-601.
14. Mubaraq MF, Maharani W. Sentiment Analysis on Twitter Social Media towards Climate Change on Indonesia Using IndoBERT Model. JURNAL MEDIA INFORMATIKA BUDIDARMA. 2022 Oct;6(4):2426.
15. Koto F, Lau JH, Baldwin T. IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization; 2021.
16. Saadah S, Auditama KM, Fattahila AA, Amorokhman FI, Aditsania A, Rohmawati AA. Implementation of BERT, IndoBERT, and CNN-LSTM in Classifying Public Opinion about COVID-19 Vaccine in Indonesia. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi). 2022 Aug;6(4):648-55.
17. et al BW. IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding; unknown. https://github.com/annisanurulazhar/absa-playground.
18. Lee D, Park J, Shim J, Lee SG. An Efficient Similarity Join Algorithm with Cosine Similarity Predicate; 2010. .

Table 8. Contoh Hasil Identifikasi yang Tidak Sesuai

No.	Topik Pertanyaan	Klaster Pertanyaan (Dataset)	Klaster Pertanyaan (Hasil Similarity Check)	Similarity Score	Status
1	apa saja obat untuk covid-19	Obat atau vitamin apa saja yang dapat digunakan untuk membantu pemulihan COVID-19 ?	Apa saja jenis vaksin COVID-19 ?	0.856033921	Tidak Sama
2	apakah pemerintah telah melakukan penilitian mengenai obat covid ?	Obat atau vitamin apa saja yang dapat digunakan untuk membantu pemulihan COVID-19 ?	Apakah vaksin dibutuhkan untuk menagani COVID-19?	0.689313889	Tidak Sama
3	apakah batuk hampir 2 minggu merupakan gejala covid-19	Apa gejala dari COVID-19?	Berapa lama gejala COVID-19 hilang?	0.801061988	Tidak Sama
4	apakah penyintas kurang dari 3 bulan sudah bisa ikut program vaksin?	Kapan diperbolehkan vaksin?	Apakah penyitas COVID-19 harus menunggu sampai 3 bulan setelah dinyatakan sembuh untuk bisa mendapatkan vaksin?	0.708156526	Tidak Sama