1. Pendahuluan

Latar Belakang

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 Covid-19 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 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 *(fine-tuned)* 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.

Topik dan Batasannya

Topik dan batasan masalah yang digunakan dalam penelitian tugas akhir ini mencakup bagaimana mengimplementasikan dan menganalisis identifikasi *similar question* dengan memanfaatkan representasi dari model IndoBERT terhadap data pertanyaan terkait Covid-19, dan bagaimana mengetahui seberapa akurat model IndoBERT dalam mengidentifikasi *similar question* berdasarkan kesesuaian dari perbandingan pertanyaan hasil identifikasi dengan pertanyaan pada dataset. Dalam penelitian ini, dataset yang digunakan adalah data kumpulan pertanyaan seputar Covid-19 yang diperoleh dari berbagai macam sumber yang totalnya berjumlah 725 pertanyaan dan beberapa nama kolom. Jenis jenis nama kolom pada dataset ini adalah sebagai berikut.

Tabel 1. Jenis label pada *dataset*

Nama Kolom	Keterangan
Teks Pertanyaan	Pertanyaan dari beberapa user

Topik Pertanyaan	Teks pertanyaan yang dibuat agar lebih mudah dipahami
Klaster Pertanyaan	Topik pertanyaan yang telah dirangkai menjadi kalimat tanya yang lebih sempurna
Kategori	Poin utama dari pertanyaan
Jawaban	Jawaban dari setiap pertanyaan yang ada
Link	Tautan untuk sumber pertanyaan yang didapat

Tujuan

Tujuan yang ingin dicapai dalam penelitian tugas akhir ini adalah untuk memvalidasi pelabelan pada data *similar question* yang dilakukan pada penelitian sebelumnya oleh M. Z. Aonillah [9], dengan memanfaatkan representasi dari model IndoBERT terhadap data pertanyaan terkait Covid-19, dan mengetahui seberapa akurat model IndoBERT dalam mengidetifikasi *similar question* berdasarkan kesesuaian dari perbandingan pertanyaan hasil identifikasi dengan pertanyaan pada dataset.

Organisasi Tulisan

Pada bab 2 dibahas studi terkait penelitian yang dilakukan, bab 3 dibahas sistem yang dibangun, bab 4 dibahas evaluasi dari model, dan bab 5 dibahas kesimpulan dari penelitian yang dilakukan.