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

Rifki Adi Pramana[1], Ade Romadhony[2]

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

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.**

Kata kunci : pemrosesan bahasa alami, question similarity, IndoBERT, Covid-19