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

Abstract

Question Answering Systems (QAS) is a task in the field of informatics, particularly within the domain of Natural Language Processing (NLP). A QAS autonomously furnishes responses based on user-generated queries. Among the stages encompassing QAS processing, lies the task of identifying questions that bear resemblance to each other, known as similar question identification. This process aims to discern questions sharing common traits, thereby facilitating accurate response generation. In this research endeavor, we focus on the identification of similar questions within a dataset centered around Covid-19 inquiries. Our methodology involves the utilization of the IndoBERT model, leveraging cosine similarity to gauge question resemblance. Through a series of empirical experiments, we successfully identify 195 out of 611 question pairs that exhibit significant similarities. Our analysis of the identification outcomes underscores various factors influencing question similarity. These factors include sentence length, initial phrase structure, and the semantic relevance among words, which collectively contribute to the formation of similar questions.

Keywords : natural language processing, question similarity, IndoBERT, Covid-19