ABSTRACT

ENHANCING ABSTRACTIVE TEXT SUMMARIZATION OF TWITTER DATA THROUGH A HYBRID SYSTEM OF SEMANTIC SEARCH AND TRANSFORMER MODELS

By

Alma Mahira Lazuardani

19/438440/PA/18898

This paper discusses the development of abstractive text summarization models for low-resource languages, specifically Indonesian. In context of social media summarization, Transformer models have achieved state-of-the-art performance for English and other lenient languages. However, there are still limited research and scarcity of resource in Indonesian language in order to achieve good result in this task. Moreover, social media language still presented as a challenge for NLP tasks due to its noisy nature and diverse structure.

This project aims to bridge the gap between Indonesian and other highly-researched languages by introducing a hybrid system of Transformer models and noise removal techniques. This project focuses on summarizing Twitter threads in Indonesian while providing comprehensive evaluation of models using ROUGE metric and human judgment. Results show that there are 7% increase in ROUGE scores and 16.12% increase in human evaluation when evaluating the quality of summaries using the hybrid system. Moreover, this project deployed a publicly available Twitter thread summarization website as an end product.

Keywords: abstractive summarization, Neural Networks, Transformers, T5, Pegasus, BART