ABSTRACT

ANUGRAH AKBAR PRARAMADHAN, 10118969

CAUSAL LANGUAGE MODELING APPLICATION WITH *GPT-2* IN SELECTING INDONESIAN DICTIONARIES USING *PYTHON* PROGRAMMING LANGUAGES.

Undergraduate Thesis, Information System, Faculty of Computer Science and Information Technology, Gunadarma University, 2023.
Keywords: Benchmark Dataset *Indo4B*, Bilingual Language Understudy (BLEU), Causal Language Modeling, GPT-2, Python Programming Language.
(xvi + 94 + Appendix)

Generative Pre-Trained Transformer 2 (GPT-2) size Medium is one of the large language model architecture developed by OpenAI in 2019 with a total of 345 million parameters. The transformer-based model is famous for the ability of unsupervised learning from an unstructured data in recognizing data patterns quite well. This study aims to create web-based applications using the Python programming language in solving the objectives of causal language modeling against the selection of Indonesian dictionaries trained using GPT-2 size Medium on the *Indo4B* benchmark dataset. The study uses the waterfall method in application development, consisting of stages of requirement, analysis, design, development, testing, deployment, and maintenance. Based on the test results, this study obtained the Bilingual Language Understudy (BLEU) value against the model of the test object A with the mean value in percent of *66.16*. The study managed to maintain the BLEU value at threshold *>50* as a reference that this model is able in producing phasing sentences and able in choosing diction of the Indonesian language very well in accordance with the principles of the Great Dictionary of Indonesia Language (KBBI). Due to its training using *Indo4B* benchmark dataset, this study succeeded in performing the adaptation of the ability similar to the previous research. This research model is capable of producing partially journalistic articles, providing informal and formal sentence information, producing food-making steps, continuing a story, explain a general term definition, and question-answering with notes still in the objective of causal language modeling that will proceed from an input sentence.

Bibliography (2017-2023)