

IndoBERT for University Chatbot Using Intent Classification

Ni Gusti Ayu Putu Meyrasinta Susila¹, Anindya Aprilianti Pravitasari*², Afrida Helen³, Najma Rafifah Putri Syallya⁴

^{1, 2, 4} Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, INDONESIA.

³ Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, INDONESIA.

(E-mail: gusti20002@mail.unpad.ac.id, anindya.aprilianti@unpad.ac.id, helen@unpad.ac.id, najma20005@mail.unpad.ac.id)

Keywords: Chatbot, IndoBERT, Intent classification.

ABSTRACT

Educational institutions commonly maintain websites to disseminate information to a broad spectrum of stakeholders, ranging from students to the general public. Using a website can be beneficial, but the website is less interactive. With technological advances; especially in the field of Natural Language Processing (NLP), websites can be enriched with chatbots capable of promptly addressing user inquiries. One frequent topic of interest, especially among graduate students, pertains to their theses or final projects. The inclusion of a chatbot enables students to easily and swiftly access details related to thesis preparation.

The objective of this study is to develop a model that can be further refined into a chatbot using the information retrieval approach with intent classification. The data used for this research was collected through a questionnaire distributed to all students in the Department of Statistics at Universitas Padjadjaran. The data gathered comprised questions that users might pose to the chatbot regarding their theses.

The method employed for intent classification involves the utilization of Indonesian Bidirectional Encoder Representations from Transformers (IndoBERT) [1]. IndoBERT is a variant of BERT that has been trained using Indonesian-language data. BERT, as part of the Transformers family, processes words in sentences based on their contextual relationships within the entire sentence [2]. BERT can be employed for various tasks, including paraphrasing, entailments, question answering, text classification, and sequence tagging [3]. BERT operates differently from other algorithms in that it comprehends words by considering their contextual significance within the given context.

Several experiments were conducted to fine-tune the IndoBERT model for intent classification. With just 10 training epochs, the accuracy in processing the data reached 95.97%. This outcome implies that the IndoBERT model is capable of effectively

capturing data patterns for intent classification, even with a relatively small number of training epochs.

Acknowledgment: The authors are grateful to the Research Center for AI and Big Data (AIDA) Universitas Padjadjaran and the Department of Statistics Universitas Padjadjaran which supports this research.

References:

- [1] Koto, F., Rahimi, A., Lau, J. H., and Baldwin, T.: ‘IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP’. 2020. <http://arxiv.org/abs/2011.00677>
- [2] Chen, Q., Zhuo, Z., and Wang, W.: ‘BERT for Joint Intent Classification and Slot Filling’. 2019. <http://arxiv.org/abs/1902.10909>
- [3] Devlin, J., Chang, MW., Lee, K., et al.: 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', Proceedings of NAACL-HLT. 2019. Pp. 4171–4186.