

INTISARI

Perkembangan bidang pemrosesan bahasa alami mengalami peningkatan pesat dalam beberapa dekade terakhir. Inovasi seperti Transformer yang paling mutakhir membuat permasalahan-permasalahan yang berkaitan dengan pemodelan bahasa seperti pengambilan informasi dalam teks, klasifikasi, tanya jawab, hingga identifikasi sentimen dapat dilakukan. Hal ini tidak luput pula mempengaruhi sektor perbankan yang memanfaatkan inovasi tersebut untuk mengatasi permasalahan yang dihadapi. Berawal dari pengembangan sistem bot pesan untuk mempermudah layanan pelanggan dengan mengimplementasikan teknologi awal NLP, muncul permasalahan ketika dihadapkan dengan bentuk bahasa yang tidak baku. Permasalahan tersebut diharapkan dapat teratasi dengan melihat penelitian NLP terbaru, yaitu menggunakan model bahasa berbasis Transformer.

Penelitian ini mengkaji bagaimana penggunaan Transformer, khususnya yang berarsitektur Llama 2 7B, untuk menyelesaikan permasalahan klasifikasi intensi pesan pengguna perbankan yang tidak bisa dijawab oleh sistem yang sudah ada. Pendekatan yang dilakukan adalah dengan mengeksplorasi metode teknik *prompt engineering* dan perlakuan *fine-tuning* QLoRA. Hasil dari pengujian menunjukkan perlakuan *prompt engineering* pada model *baseline* didapat teknik *few-shot* + *CoT* untuk *prompt* Bahasa Inggris dan teknik *few-shot* + *CoT* + *bothinst* + *mock* + *reit* + *strict* + *info* + *name* untuk *prompt* Bahasa Indonesia. Keduanya memiliki performa skor F1 terbaik sebesar 0,78 dan 0,83. Di sisi lain, ketika dilakukan pengujian teknik *prompt engineering* terbaik pada model terbuka, Merak-v3 memiliki performa terbaik pada kasus kedua bahasa *prompt* yang digunakan. Model tersebut memiliki nilai skor F1 0,46 dan akurasi 0,35 pada Bahasa Indonesia, serta skor F1 0,54 dan akurasi 0,50 pada Bahasa Inggris. Pada kasus yang berbeda, penerapan *fine-tuning* QLoRA pada model bahasa terbuka membuktikan bahwa teknik *fine-tuning* tersebut secara signifikan meningkatkan performa setiap model terbuka yang dipilih. Lebih lanjut, model Cendol berbasis Llama-2-7B memiliki hasil terbaik pada kasus *fine-tuning* ini dengan nilai skor F1 dan akurasi masing-masing sebesar 0,93. Hal ini menunjukkan bahwa terdapat potensi penerapan pendekatan yang dikembangkan tersebut untuk diimplementasikan ke dalam sistem yang sudah ada agar dapat membantu deteksi intensi pesan yang tidak bisa dijawab oleh sistem tersebut.

Kata kunci : klasifikasi intensi, perbankan, transformer, *prompt engineering*, *fine-tuning* QLoRA

ABSTRACT

The field of natural language processing has seen rapid growth in recent decades. Innovations such as the latest Transformer models have enabled the handling of various language-related tasks, such as information retrieval in texts, classification, question answering, and sentiment identification. This progress has also impacted the banking sector, which has leveraged these innovations to address its challenges. The development of chatbot systems to facilitate customer service by implementing early NLP technology initially faced issues when dealing with non-standard language forms. It is hoped that these issues can be addressed by examining the latest NLP research, particularly through the use of transformers.

This study examines the use of Transformers, particularly those with Llama 2 7B architecture, to solve the problem of intent classification for banking user messages that cannot be handled by existing systems. The approach involves exploring the use of prompt engineering techniques and applying QLoRA fine-tuning. The results of the experiments show that baseline model with prompt engineering using the "*few-shot + CoT*" technique for English prompts and the "*few-shot + CoT + bothinst + mock + reit + strict + info + name*" technique for Indonesian prompts achieved the best F1 scores of 0.78 and 0.83, respectively. Meanwhile, testing on an open-source model with those best prompt engineering techniques revealed that the Merak-v3 model performed best on both language prompt cases, with F1 scores of 0.46 and an accuracy of 0.35 for Indonesian, and an F1 score of 0.54 and an accuracy of 0.50 for English. Applying QLoRA fine-tuning to the selected open-source language models demonstrated that this fine-tuning technique significantly improved the performance of each chosen open-source model. The best fine-tuning results were achieved by the Cendol model based on Llama-2-7B, with F1 score and accuracy values of 0.93. This indicates the potential for implementing the developed approach into existing systems to help detect user message intents that cannot be answered by these systems.

Keywords: intent classification, banking, transformer, prompt engineering, fine-tuning QLoRA