PERBANDINGAN MODEL GPT-3.5 TURBO DAN LLAMA-2 DALAM ANALISIS SENTIMEN ULASAN PADA BLUE KARMA UBUD

Oleh

I Wayan Adi Maha Wiguna, NIM 2215051049

Program Studi Pendidikan Teknik Informatika

Jurusan Teknik Informatika

ABSTRAK

Large Language Models (LLM) memerlukan metode tambahan untuk optimasi pada tugas spesifik seperti analisis sentimen. Penelitian ini membandingkan performa GPT-3.5 Turbo dan LLaMA-2 melalui penerapan metode *Retrieval Augmented Few-shot* (RAFS) pada domain pariwisata, dengan skenario *Zero-shot* sebagai *baseline*. Hasil eksperimen menunjukkan bahwa LLaMA-2 mengalami peningkatan performa yang jauh lebih signifikan dibandingkan GPT-3.5 Turbo setelah penerapan RAFS. Pada proporsi data 100% akurasi LLaMA-2 meningkat dari 0,833 menjadi 0,871, sementara GPT-3.5 Turbo hanya meningkat tipis dari 0,851 menjadi 0,854. Pada proporsi dataset 50% menunjukan peningkatan yang signifikan pada kedua model, yang dimana LLaMA-2 mendapat akurasi 0.880 sedangkan GPT 3.5 Turbo mendapat akurasi 0.868, yang dimana menunjukan hasil yang lebih baik. Secara *head-to-head*, LLaMA-2 terbukti sedikit lebih unggul dibanding dengan GPT-3.5 Turbo dalam menghasilkan klasifikasi yang tepat dan seimbang. Meskipun GPT-3.5 Turbo memiliki *baseline* awal yang lebih tinggi, LLaMA-2 menunjukkan kemampuan adaptasi dan skalabilitas yang lebih baik terhadap augmentasi konteks.

Kata Kunci: Analisis Sentimen, GPT, LLaMA, Retrieval Augmented Few-shot

COMPARISON OF GPT-3.5 TURBO AND LLAMA-2 MODELS IN SENTIMENT ANALYSIS OF REVIEWS AT BLUE KARMA UBUD

By

I Wayan Adi Maha Wiguna, NIM 2215051049

Informatics Engineering Education Study Program

Department of Informatics Engineering

ABSTRACT

Large Language Models (LLMs) require additional methods to optimize performance for specific tasks such as sentiment analysis. This study compares the performance of GPT-3.5 Turbo and LLaMA-2 through the application of the Retrieval Augmented Few-shot (RAFS) method in the tourism domain, with a Zero-shot scenario used as the baseline. The experimental results show that LLaMA-2 experiences a much more significant performance improvement compared to GPT-3.5 Turbo after the implementation of RAFS. With 100% data proportion, the accuracy of LLaMA-2 increased from 0.833 to 0.871, while GPT-3.5 Turbo only slightly improved from 0.851 to 0.854. With a 50% dataset proportion, both models showed significant improvement, where LLaMA-2 achieved an accuracy of 0.880 while GPT-3.5 Turbo reached 0.868, indicating better results overall. In a head-to-head comparison, LLaMA-2 proved to be slightly superior to GPT-3.5 Turbo in producing accurate and balanced classifications. Although GPT-3.5 has a higher initial baseline, LLaMA-2 demonstrates better adaptability and scalability when contextual augmentation is applied.

Keywords: Sentiment Analysis, GPT, LLaMA, Retrieval Augmented Few-shot.