ABSTRAK

Seiring dengan meningkatnya jumlah dan popularitas portal berita digital, kini sudah lazim menjumpai iklan native yang disajikan dalam format mirip artikel. Iklan native berisiko menipu pembaca sehingga mengira iklan tersebut sebagai artikel asli. Oleh karena itu, terdapat kebutuhan yang semakin meningkat untuk mengembangkan sistem klasifikasi otomatis. Salah satu model bahasa mutakhir yang unggul dalam klasifikasi teks adalah IndoBERT, meskipun model ini terbatas pada pemrosesan urutan hingga 512 token, sedangkan dokumen berita berbentuk panjang dapat melampaui batas tersebut. Penelitian ini mengkaji kinerja model bahasa IndoBERT dalam klasifikasi teks panjang, dengan membandingkan dua pendekatan: Skenario A (*baseline* dengan teknik *truncation*) dan Skenario B (sliding window dengan attention pooling). Hasil eksperimen menunjukkan bahwa penggunaan sliding window memberikan peningkatan metrik macro-F1 sebesar 0,0026 dibandingkan baseline. Namun, analisis statistik menunjukkan bahwa perbedaan tersebut tidak signifikan secara nyata ($p > 0{,}05$) dengan effect size yang tergolong kecil. Di sisi lain, Skenario A terbukti jauh lebih efisien dengan selisih waktu komputasi lebih cepat dibandingkan Skenario B. Temuan ini menyimpulkan bahwa untuk tugas klasifikasi ini, pendekatan truncation standar pada IndoBERT merupakan strategi yang lebih optimal karena memberikan keseimbangan performa dan efisiensi sumber daya yang lebih baik dibandingkan penambahan kompleksitas melalui sliding window.

ABSTRACT

With the growing number and popularity of digital news portals, it has become common to encounter native ads presented in a format similar to articles. Native ads risk misleading readers into thinking they are genuine articles. Therefore, there is a growing need to develop automated classification systems. One of the state-of-the-art language models that excels in text classification is IndoBERT; however, this model is limited to processing sequences of up to 512 tokens, whereas long-form news articles can exceed this limit. This study examines the performance of the IndoBERT language model in classifying long texts by comparing two approaches: Scenario A (baseline with truncation) and Scenario B (sliding window with attention pooling). The experimental results show that using a sliding window yields a 0.0026 improvement in the macro-F1 metric compared to the baseline. However, statistical analysis indicates that this difference is not statistically significant ($p > 0.05$) and has a relatively small effect size. On the other hand, Scenario A proved to be far more efficient, with a significantly shorter computation time compared to Scenario B. These findings conclude that for this classification task, the standard truncation approach in IndoBERT is a more optimal strategy because it provides a better balance of performance and resource efficiency compared to the added complexity introduced by the sliding window.