

ABSTRACT

ENHANCING CLIMATEBERT USING LSG ATTENTION AND WEIGHTED CROSS ENTROPY LOSS TO CLASSIFY LONG AND IMBALANCED CLIMATE DATA

Maeve Zahwa Adriana Crown Zaki

21/477527/PA/20675

Climate change is one of the most pressing global issues that requires in-depth text data analysis. However, in processing text-based climate data, problems such as word length and class imbalance often hinder model performance. Therefore, this research modifies ClimateBERT using LSG (Local, Sparse, Global) attention and weighted cross-entropy loss to address both of these problems using a climate-based model. From experiments conducted on three datasets with different characteristics, it is known that the Weighted-LSG-ClimateBERT model succeeded in becoming the best model on two of the three datasets with an increase in F1 score ranging from 6-8%, although the improvement cannot be said to be significant because it is not consistent across all classes in the dataset. This indicates that the characteristics of the dataset influence the most suitable model architecture to be used, and that the weighted loss function is quite influential in improving classification performance on imbalanced datasets, although not significantly. Further analysis shows that LSG attention provides computational advantages and better performance compared to Longformer, not only on long texts but also on short texts.

Keywords: ClimateBERT, LSG Attention, Weighted Cross-Entropy Loss, Text Classification, Climate Data, Imbalance Dataset, Long Da

INTISARI

MODIFIKASI CLIMATEBERT MENGGUNAKAN LSG ATTENTION DAN WEIGHTED CROSS-ENTROPY LOSS UNTUK KLASIFIKASI DATA IKLIM BERUKURAN PANJANG DAN TIDAK SEIMBANG

Maeve Zahwa Adriana Crown Zaki

21/477527/PA/20675

Perubahan iklim merupakan salah satu isu global paling mendesak yang memerlukan analisis data teks secara mendalam. Namun, dalam memproses data iklim berbentuk teks, permasalahan seperti panjang kata serta ketidakseimbangan kelas sering kali menghambat performa model. Oleh karena itu, penelitian ini memodifikasi ClimateBERT menggunakan LSG (*Local, Sparse, Global*) attention dan *weighted cross-entropy loss* untuk mengatasi kedua permasalahan tersebut dengan menggunakan model berbasis *climate*. Dari eksperimen yang dilakukan pada tiga *dataset* dengan karakteristik berbeda, diketahui bahwasanya model Weighted-LSG-ClimateBERT berhasil menjadi model terbaik pada dua dari tiga *dataset* dengan peningkatan skor F1 berkisar antara 6-8%, meskipun peningkatannya tidak dapat dikatakan signifikan karena tidak konsisten di seluruh kelas dalam *dataset*. Hal ini menunjukkan bahwasanya karakteristik *dataset* berpengaruh terhadap arsitektur model yang paling cocok digunakan, serta bahwa fungsi *weighted loss* cukup berpengaruh untuk meningkatkan performa klasifikasi pada *dataset* yang tidak seimbang, meskipun tidak signifikan. Analisis lebih lanjut menunjukkan bahwa LSG attention memberikan keuntungan komputasi dan kinerja yang lebih baik dibandingkan Longformer tidak hanya pada teks yang panjang, tetapi juga pada teks pendek.

Kata Kunci: ClimateBERT, LSG Attention, Weighted Cross-Entropy Loss, Klasifikasi Teks, Data Iklim, Ketidakseimbangan Kelas, Data Panjang