ABSTRACT

This study aims to explore the use of the XLM-RoBERTa model for emotion classification on bilingual text, specifically in Indonesian and English. It examines the effect of emoji representation and epoch settings as part of hyperparameter exploration. Before training, text was cleaned from noise such as special characters. Two training scenarios were used: converting emojis into descriptive text and removing them, with three epoch settings: 3, 5, and 10. Early stopping was applied to prevent overfitting. Evaluation was carried out using accuracy and confusion matrix for each scenario. Results show that the presence of emoji positively contributes to model accuracy, achieving a maximum of 96.33% for Indonesian text and 93.45% for English. Moreover, the optimal number of epochs varies across scenarios. The English dataset shows different outcomes depending on emoji presence, while the Indonesian dataset remains consistent with a single configuration. These findings indicate that data complexity and characteristics influence model training. This research contributes to the development of emotion classification models for bilingual text by considering nonverbal elements such as emojis.

Keywords — Bilingual Text, Emoji, Hyperparameter Exploration, Noise Character, XLM-RoBERTa

ABSTRAK

Penelitian ini bertujuan mengeksplorasi penerapan model *XLM-RoBERTa* dalam klasifikasi emosi berbasis teks *bilingual*, khususnya Bahasa Indonesia dan Bahasa Inggris. Fokus utama terletak pada analisis pengaruh representasi emoji dalam teks serta konfigurasi jumlah *epoch* sebagai bagian dari eksplorasi *hyperparameter*. Sebelum pelatihan, teks dibersihkan dari karakter noise seperti simbol khusus. Penelitian dilakukan melalui dua skenario pelatihan, yaitu mengonversi emoji menjadi teks deskriptif, dan menghapusnya, dengan tiga konfigurasi *epoch*: 3, 5, dan 10. Teknik *early stopping* digunakan untuk mencegah *overfitting*. Evaluasi dilakukan menggunakan akurasi dan *confusion matrix* pada masing-masing skenario. Hasil menunjukkan bahwa keberadaan emoji berkontribusi positif terhadap akurasi model, dengan akurasi tertinggi sebesar 96.33% pada data Bahasa Indonesia dan 93.45% pada data Bahasa Inggris. Selain itu, konfigurasi *epoch* optimal tidak selalu sama pada tiap skenario. Data Bahasa Inggris menunjukkan perbedaan hasil tergantung keberadaan emoji, sedangkan data Bahasa Indonesia cenderung konsisten pada satu konfigurasi. Hal ini mengindikasikan bahwa kompleksitas serta karakteristik data memengaruhi proses pelatihan model. Penelitian ini berkontribusi pada pengembangan model klasifikasi emosi pada teks *bilingual* dengan mempertimbangkan unsur nonverbal seperti emoji.

Kata Kunci — Eksplorasi *Hyperparameter*, Emoji, *Noise Character*, Teks *Bilingual*, *XLM-RoBERTa*