INTISARI

AUTHORSHIP VERIFICATION PADA *TWEET* BERBAHASA INDONESIA DENGAN MENGGUNAKAN *EMOJI2VEC* DAN *EMOJI2TEXT*

Oleh

Yumna Aqila Rasyidah

20/462195/PA/20167

Authorship verification adalah proses untuk memverifikasi identitas penulis dari suatu teks atau dokumen digital, dengan tujuan utama memastikan keaslian dan keotentikan karya tulis melalui identifikasi pola gaya penulisan khas seorang penulis. Proses ini penting untuk menentukan apakah dua teks ditulis oleh penulis yang sama atau tidak. Penelitian ini mengeksplorasi penggunaan model IndoBERT dengan tambahan fitur emoji untuk meningkatkan *authorship verification* pada teks Twitter berbahasa Indonesia. Dua pendekatan representasi emoji yang diuji adalah *emoji2vec* dan *emoji2text*. *Emoji2vec* adalah representasi vektor kata *pretrained* yang digunakan untuk semua emoji Unicode, sementara *emoji2text* adalah *dataset* deskripsi emoji dalam bahasa Indonesia.

Penggunaan emoji sebagai fitur tambahan dalam *authorship verification* terbukti meningkatkan kinerja model secara signifikan. Hasil penelitian menunjukkan bahwa metode *emoji2vec* + IndoBERT memberikan performa terbaik dengan akurasi 88,12%, presisi 91,89%, *recall* 83,63%, F1 *score* 87,56%, dan ROC-AUC *score* 0,88, dibandingkan dengan metode tanpa emoji dan *emoji2text* yang memiliki performa lebih rendah. Walaupun waktu prediksi pada pengujian *emoji2vec* sedikit lebih lama (8,27 detik) dibandingkan pengujian lainnya, peningkatan metrik performa lainnya menjadikan metode ini sebagai pilihan terbaik untuk *authorship verification* pada teks Twitter berbahasa Indonesia.

Kata Kunci: *Authorship Verification, Emoji2vec, Emoji2Text, Natural Language Processing*

ABSTRACT

AUTHORSHIP VERIFICATION ON INDONESIAN TWEETS USING EMOJI2VEC AND EMOJI2TEXT

By

Yumna Aqila Rasyidah

20/462195/PA/20167

Authorship verification is the process of verifying the identity of the author of a text or digital document, with the primary goal of ensuring the authenticity and originality of the written work by identifying the unique writing style of an author. This process is essential to determine whether two texts presented to the system were written by the same author or not. This study explores the use of the IndoBERT model with the addition of emoji features to enhance the performance of authorship verification on Indonesian Twitter texts. Two emoji representation approaches tested are emoji2vec and emoji2text. Emoji2vec is a pretrained word vector representation used for all Unicode emojis, while emoji2text is a *dataset* of emoji descriptions in Indonesian.

The use of emojis as an additional feature in authorship verification has proven to significantly improve model performance. The results show that the emoji2vec + IndoBERT method provides the best performance with an accuracy of 88.12%, precision of 91.89%, recall of 83.63%, F1 score of 87.56%, and ROC-AUC score of 0.88, compared to methods without emojis and emoji2text which have lower performance. Although the prediction time in the emoji2vec test is slightly longer (8.27 seconds) compared to other tests, the improvement in other performance metrics makes this method the best choice for authorship verification on Indonesian Twitter texts.

Keywords: Authorship Verification, Emoji2vec, Emoji2Text, Natural Language Processing