ABSTRAK

Depresi merupakan gangguan kesehatan mental dengan prevalensi tinggi yang berdampak luas pada individu dan masyarakat. Meskipun banyak individu mengekspresikan kondisi psikologisnya melalui media sosial X (Twitter), deteksi dini masih terbatas karena stigma sosial dan rendahnya akses layanan kesehatan mental. Penelitian terdahulu pada bidang ini lebih berfokus pada peningkatan akurasi model tanpa menjelaskan pola bahasa yang mendasari hasil prediksi, sehingga model bersifat black box dan sulit dipercaya dalam konteks kesehatan mental yang sensitif.

Penelitian ini menerapkan model pra-latih IndoBERTweet yang dikombinasikan dengan metode *Explainable Artificial Intelligence* (XAI) berbasis *Shapley Additive Explanations* (SHAP) untuk mendeteksi dan menjelaskan prediksi depresi pada tweet berbahasa Indonesia. Data dikumpulkan melalui web scraping sebanyak 25.732 tweet dari platform X, kemudian diproses melalui *preprocessing*, pelabelan otomatis berbasis model emosi IndoBERT, dan fine-tuning IndoBERTweet untuk klasifikasi biner depresi dan normal.

Hasil evaluasi menunjukkan model mencapai akurasi 94,43% dengan weighted F1-Score 94,44%. Evaluasi correctness menggunakan metrik AOPC membuktikan token yang diidentifikasi SHAP benar-benar berpengaruh terhadap keputusan model dengan gap tertinggi sebesar 0,4051 pada K=12. Evaluasi coherence menunjukkan 12 dari 20 token SHAP teratas sangat sesuai dengan indikator klinis DSM-5, seperti berduka, hampa, sedih, menyesal, bersalah, kesepian, gagal, lelah, kehilangan, dan kecewa. Penelitian ini membuktikan bahwa SHAP mampu menghasilkan penjelasan yang valid secara statistik dan koheren dengan pengetahuan domain psikologi.

Kata Kunci: Depresi, IndoBERTweet, *Explainable Artificial Intelligence*, SHAP, NLP

ABSTRACT

Depression is a mental health disorder with high prevalence that broadly impacts individuals and society. Although many individuals express their psychological state through social media platform X (Twitter), early detection remains limited due to social stigma and inadequate access to mental health services. Prior research in this field has focused primarily on improving model accuracy without explaining the linguistic patterns underlying prediction outcomes, rendering models black-box in nature and difficult to trust within the sensitive context of mental health.

This study applies the pre-trained IndoBERTweet model combined with an Explainable Artificial Intelligence (XAI) method based on Shapley Additive Explanations (SHAP) to detect and explain depression predictions in Indonesian-language tweets. Data were collected via web scraping, yielding 25,732 tweets from platform X, which were then processed through preprocessing, automatic labeling using an IndoBERT-based emotion model, and fine-tuning of IndoBERTweet for binary classification of depression and normal.

Evaluation results show the model achieved an accuracy of 94.43% with a weighted F1-Score of 94.44%. Correctness evaluation using the AOPC metric demonstrates that tokens identified by SHAP genuinely influence the model's decisions, with the highest gap of 0.4051 at K=12. Coherence evaluation reveals that 12 out of 20 top SHAP tokens are highly consistent with DSM-5 clinical indicators, including berduka (grief), hampa (emptiness), sedih (sadness), menyesal (regret), bersalah (guilt), kesepian (loneliness), gagal (failure), lelah (fatigue), kehilangan (loss), and kecewa (disappointment). This study demonstrates that SHAP is capable of producing explanations that are both statistically valid and coherent with domain knowledge in psychology.

Keywords: *Depression, IndoBERTweet, Explainable Artificial Intelligence, SHAP, NLP*