

INTISARI

Klasifikasi *Trait* Kepribadian Multi-label pada Teks Bahasa Indonesia di Media Sosial Menggunakan Fine-Tuning BERT

Oleh

AZARI BERESMAN

23/529231/PPA/06695

Perkembangan media sosial menghasilkan data teks dalam jumlah besar yang berpotensi dimanfaatkan untuk analisis psikologis, termasuk prediksi kepribadian berbasis Big Five (OCEAN). Namun, sebagian besar penelitian sebelumnya masih berfokus pada bahasa Inggris dan memformulasikan prediksi sebagai klasifikasi *single-label* atau *binary per trait*, sehingga belum sepenuhnya merepresentasikan sifat multidimensional kepribadian. Pada konteks bahasa Indonesia, keterbatasan dataset berlabel tervalidasi serta variasi bahasa informal media sosial menjadi tantangan tersendiri.

Penelitian ini bertujuan untuk memformulasikan prediksi kepribadian Big Five sebagai permasalahan *multi-label classification* berbasis teks Bahasa Indonesia serta mengevaluasi efektivitas berbagai model pembelajaran mesin. Pendekatan yang digunakan meliputi *Logistic Regression*, *Support Vector Machine*, *Extreme Gradient Boosting*, dan *fine-tuning* IndoBERT. Untuk meningkatkan robustnes model terhadap variasi linguistik, diterapkan juga *data augmentation* dengan beberapa jenis strategi yaitu non-kontekstual, semi-kontekstual, dan kontekstual.

Eksperimen dilakukan menggunakan dataset *stream-of-consciousness essays* berlabel psikometrik sebagai data latih dan data tweet pengguna yang telah mengisi tes Big Five sebagai data uji eksternal. Evaluasi menggunakan metrik *Precision*, *Recall*, dan F1-score pada skema multi-label. Hasil menunjukkan bahwa model berbasis transformer dengan *data augmentasi* memberikan performa terbaik dengan *precision* 0.73, *recall* 0.95, dan *F1-Score* 0.83, melampaui model klasik. Penerapan *data augmentation* juga terbukti mampu meningkatkan performa setiap model dengan nilai sebesar 0.05 – 0.08 pada F1-score.

Penelitian ini menunjukkan bahwa formulasi multi-label yang selaras dengan teori psikologi, dikombinasikan dengan representasi kontekstual dan augmentasi terstruktur, dapat meningkatkan efektivitas model untuk prediksi kepribadian pada teks media sosial Bahasa Indonesia.

Kata Kunci : Big-Five Personality, Klasifikasi Multi-label, *Fine-tuning* IndoBERT, *Data Augmentation*

ABSTRACT

Multi-label Personality Trait Classification in Indonesian Language Text on Social Media using Fine-Tuning BERT

by

AZARI BERESMAN

23/529231/PPA/06695

The rapid growth of social media has generated a vast amount of textual data that can be leveraged for psychological analysis, including personality prediction based on the Big Five Personality Traits (OCEAN). However, most previous studies have focused primarily on English-language data and have formulated personality prediction as single-label or binary classification per trait, which does not fully capture the multidimensional nature of personality. In the context of the Indonesian language, the limited availability of psychometrically validated labeled datasets and the high variability of informal social media language pose additional challenges.

This study aims to formulate Big Five personality prediction as a multi-label classification problem using Indonesian text and to evaluate the effectiveness of various machine learning approaches. The proposed models include Logistic Regression, Support Vector Machine, Extreme Gradient Boosting, and fine-tuned IndoBERT. To enhance model robustness against linguistic variation, several data augmentation strategies were implemented, including non-contextual, semi-contextual, and contextual approaches.

Experiments were conducted using a psychometrically labeled stream-of-consciousness essay dataset as training data and tweets from users who completed the Big Five personality test as external evaluation data. Model performance was evaluated using Precision, Recall, and F1-score under a multi-label framework. The results indicate that transformer-based models with data augmentation achieved the best performance, with a precision of 0.73, recall of 0.95, and F1-score of 0.83, outperforming classical machine learning models. Furthermore, the application of data augmentation improved model performance across all approaches, yielding F1-score gains ranging from 0.05 to 0.08.

These findings demonstrate that a multi-label formulation aligned with psychological theory, combined with contextual language representation and structured augmentation strategies, enhances the effectiveness of personality prediction models for Indonesian social media text.

Keywords : Big-Five Personality, Multi-label Classification, Fine-tuning IndoBERT, Data Augmentation