Prediksi Kepribadian *Big Five* Pengguna Twitter Menggunakan Metode *Decision Tree* dengan Pendekatan Semantik *BERT*

Jammie Reyhan Widyanto, Dr. Erwin Budi Setiawan S.Si., M.T.

Fakultas Informatika, Universitas Telkom, Bandung

jammiereyhan@students.telkomuniversity.ac.id,
erwinbudisetiawan@telkomuniversity.ac.id

Abstrak

Kepribadian individu dapat dilihat dengan mudah pada zaman ini. Ada beberapa pendekatan dalam mengklasifikasi kepribadian, salah satunya kepribadian *big five*. Kepribadian *big five* terdiri dari 5 dimensi, yaitu *Openness, Conscientiousness, Extraversion, Agreeableness,* dan *Neuroticism*. Salah satu cara mengetahui kepribadian individu dapat dilihat dari media sosialnya, karena zaman ini hampir semua individu mempunyai media sosial. Salah satu media sosial yang masih ramai digunakan adalah Twitter. Twitter adalah media sosial yang berisi cuitan-cuitan atau *tweet* dari tiap individu yang maksimal 280 karakter per-*tweet*. Sudah ada beberapa penelitian terkait kepribadian *big five* dari pengguna twitter. Berdasarkan permasalahan penelitian kepribadian *big five* sebelumnya, maka pada penelitian ini dilakukan prediksi kepribadian *big five* pengguna twitter dengan menggunakan metode *Decision Tree Classification And Regression Tree (CART), Term Frequency – Invers Document Frequency (TF-IDF), Synthetic Minority Oversampling Technique (SMOTE), Linguistic Inquiry Word Count (LIWC),* dan *Bidirectional Encoder Representations from Transformers (BERT)*. Penelitian bertujuan untuk mengetahui penerapan metode-metode yang digunakan pada penelitian ini terhadap prediksi kepribadian *big five* dan untuk mendapatkan hasil akurasi yang lebih baik dari penelitian sebelumnya. Data yang didapatkan pengguna twitter berjumlah 315 pengguna twitter dan 672.866 *tweet* yang diperoleh dari survey dan telah dilabeli kepribadian *big five*, dihasilkan akurasi sebesar 97,62% dari baseline dengan kenaikan 23.1%, dengan menerapkan metode *CART+TF-IDF+SMOTE+LIWC+BERT*
Kata kunci: *CART,* Kepribadian Big Five*, LIWC, SMOTE, BERT,* Twitter.

Abstract

Individual personality can be seen easily in this day. There are several approaches in classifying personality, one of which is the big five personality. The big five personality consists of 5 dimensions, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. One way of knowing an individual's personality can be seen from their social media, because today almost all individuals have social media. One of the social media that is still widely used is Twitter. Twitter is a social media that contains tweets from each individual with a maximum of 280 characters per tweet. There have been several studies related to the big five personalities of Twitter users. Based on previous big five personality research problems, this study carried out predictions of the big five personalities of Twitter users using the Decision Tree Classification And Regression Tree (CART), Term Frequency – Inverse Document Frequency (TF-IDF), Synthetic Minority Oversampling Technique (SMOTE), Linguistic Inquiry Word Count (LIWC), and Bidirectional Encoder Representations from Transformers (BERT) methods. The study aims to determine the application of the methods used in this study to the prediction of big five personalities and to get better accuracy results than previous studies. Data obtained from 315 twitter users and 672,866 tweets obtained from surveys and have been labeled with big five personalities, resulting in an accuracy of 97.62% from the baseline with an increase of 23.1%, by applying the CART+TF-IDF+SMOTE+LIWC+BERT method.
Keywords: CART, *Kepribadian* Big Five, LIWC, SMOTE, BERT, Twitter.

1. Pendahuluan

Kepribadian merupakan hal yang melekat pada tiap individu. Kepribadian juga mempengaruhi setiap individu pada berbagai macam kegiatan [1]. Salah satunya yaitu ada kepribadian *big five.* Kepribadian *big five* merupakan suatu pendekatan dalam psikologi untuk menentukan kepribadian dari individu dalam beberapa dimensi. Menurut Lewis R. Goldberg ada 5 dimensi yaitu, *Openness, Conscientiousness, Extraversion,*

Agreeableness, dan *Neuroticism* [2]. Salah satu cara untuk mengetahui kepribadian seseorang dapat dilihat pada media sosialnya.

Media sosial adalah suatu media yang penggunanya gunakan untuk mengekspresikan diri dalam berbagai bentuk seperti tulisan, foto, dan video sehingga akun media sosial suatu individu dapat mencerminkan penggunanya, sehingga pengguna lain dapat menyaring dengan mudah apa yang ingin dilihat atau muncul pada media sosialnya [3]. Salah satu media sosial yang masih ramai digunakan adalah Twitter. Twitter diperkenalkan sebagai media sosial yang cara mengekspresikan diri dengan mengunggah *tweet* yang maksimal karakternya hingga 280 karakter. Sehingga *tweet* pada twitter bersifat singkat dan padat.

Pada penelitian ini diusulkan deteksi kepribadian dengan model yang mendekati sempurna, metode *Classification and Regression Tree* (*CART*) yang dikombinasi dengan *TF-IDF, SMOTE, LIWC*, dan *BERT*. Metode *CART* ini digunakan pada penelitian ini karena metode *CART* memiliki ketepatan hasil dan akurasi yang baik dibandingkan dengan metode lainya [4]. Metode ini merupakan salah satu dari beberapa metode *Decision Tree* yang dikembangkan oleh Leo Breiman yang merupakan algoritma berbentuk pohon keputusan untuk mengklasifikasi, yang pada tiap parent node akan terpecah dan memiliki *child node* [5]. Metode ini juga telah digunakan sebagai metode penelitian sebelumnya tetapi dengan topik lain yang menghasilkan nilai akurasi yang baik [6]. *Term Frequency – Inverse Document Frequency* (*TF-IDF*) yang digunakan sebagai metode yang memberikan nilai atau bobot pada suatu kata berdasarkan seberapa sering kata tersebut muncul dan seberapa penting kata tersebut [7].

Metode *Synthetic Minority Oversampling Technique (SMOTE)* yang digunakan sebagai pengambilan sampel data untuk kelas minoritas dengan menghubungkan data yang dipilih secara acak dari data yang terdekat [8]. Metode *Linguistic Inquiry and Word Count (LIWC)* yang digunakan untuk menghitung kata berdasarkan kategorinya yang dikembangkan sejak tahun 2007 [9]. Metode selanjutnya yaitu menggunakan pendekatan semantic *Bidirectional Encode Representations from Transformers (BERT)* yang diperkenalkan oleh Jacob Devlin pada tahun 2018 sebagai model representasi bahasa. *BERT* digunakan sebagai pengesktrak fitur dalam teks pada *NLP* yang mempunyai dua proses yaitu, pertama dataset akan diubah menjadi vektor kalimat yang menghasilkan nilai semantik, kemudian dataset akan dibuat sebagai token yang menghasilkan pemberian label pada token, sehingga dapat diimplementasikan ke dalam klasifikasi [10].

2. Studi Terkait

Prediksi kepribadian *big five* bukan pertama kali dilakukan, sudah ada penelitian sebelumnya yang melakukannya dengan berbagai metode yang berbeda. Pada penelitian yang dilakukan Rendy, metode yang digunakan yaitu *TF-IDF* sebagai pembobot kata dan *Random Forest* sebagai metode klasifikasi, serta pada evaluasi menggunakan *confusion matrix* yang akan menghitung akurasi, *precision, recall*, dan *f1 score* [11]. Hasilnya, akurasi tertinggi pada skenario pertama yang didapatkan dari parameter standar pada perbandingan data 20:80 sebesar 57,69%. Lalu pada skenario kedua yang didapatkan dari parameter optimal pada perbandingan 20:80 sebesar 69,23%. Model dapat meningkatkan rata-rata dari akurasi sebesar 5,94% dengan peningkatan yang paling tinggi pada fitur sentiment yang perbandingannya 10:90 sebesar 23,07%. Lalu sistem mencari hubungan korelasi antara fitur dengan kepribadian dengan menggunakan *pearson correlation*.

Pada penelitian yang dilakukan Roji metode yang digunakan oleh peneliti yaitu menggunakan *TF-IDF* sebagai pembobot kata dan *K-Nearest Neighbor* sebagai metode klasifikasi, serta pada evaluasi peneliti menggunakan *confusion matrix* untuk mengukur kinerja dari metode klasifikasinya [12]. Hasilnya, akurasi tertinggi dari perilaku sosial dan linguistik dengan pengukuran performasi nilai k=9 sebesar 60,97%, sedangkan perilaku sosial dan linguistik dengan pengukuran performansi nilai k=1 yang terendah sebesar 39,02%. Dengan dilakukan item kuisoner berdasarkan jumlah yang sama pada tiap dimensi hasil skoring tetap mendapatkan label lebih banyak di salah satu dimensi yaitu *Openness,* sehingga model prediksi yang dibangun cenderung memprediksi openness dan membuat nilai *precision* dan *recall* pada openness meningkat tetapi tidak dengan dimensi lain.

Metode *CART* sebelumnya belum pernah digunakan pada studi kasus kepribadian *big five*, pada penelitian yang dilakukan Nurihsan dilakukan dengan studi kasus pasien penderita DBD. Pada penelitian tersebut menjelaskan bahwa Metode *CART* memiliki hasil klasifikasi dan akurasi yang baik, dibandingkan dengan metode lainnya. Metode *CART* digunakan untuk melakukan penelitian yang menggunakan proses klasifikasi dan regresi [13]. Pada penelitian sebelumnya metode yang digunakan peneliti yaitu *CART* dengan menggunakan data pasien penderita DBD. Pada prosesnya peneliti menggunakan data sebanyak 230 data sampel yang terbagi dua menjadi data learning sebanyak 143 sampel dan data testing sebanyak 87 sampel.