ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Klasifikasi Kepribadian Big Five Pengguna Twitter dengan Metode Naïve Bayes

Yusra[1], Muhammad Fikry[2], Rinaldi Syarfianto[3], Reski Mai Candra[4], Elvia Budianita[5]
[1,2,3,4,5] Jurusan Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau
Jl. HR. Soebrantas No. 155 Simpang Baru, Panam, Pekanbaru, 28293, 0761-562223
e-mail: yusra@uin-suska.ac.id, muhammad.fikry@uin-suska.ac.id,
rinaldi.syarfianto@students.uin-suska.ac.id, reski.candra@uin-suska.ac.id,
elvia.budianita@uin-suska.ac.id

Abstrak

Untuk dapat memahami kepribadian seseorang, postingannya di media sosial dapat digunakan sebagai sumber informasi. Pada penelitian ini, metode Naïve Bayes digunakan untuk mengklasifikasikan kepribadian pengguna Twitter ke dalam salah satu dari lima kelas, yaitu Openness, Conscientiousness, Extraversion, Agreeableness, dan Neuroticism. Tweet diunduh dari 15 akun Twitter dengan menggunakan Twitter API, dengan total keseluruhan sebanyak 1.500 tweet. Setiap akun ditetapkan sifat kepribadian dominannya berdasarkan hasil kuesioner kepribadian yang diinterpretasikan oleh seorang pakar psikologi. Setiap tweet dipraproses menjadi huruf kecil, dibersihkan, ditokenisasi menjadi kata, ditemukan kata dasarnya, kemudian dihilangkan kata-kata yang tidak penting. Setiap kata dibobot berdasarkan frekuensinya. Dataset dibagi menjadi data latih dan uji dengan perbandingan 60:40, 70:30, 80:20 dan 90:10. Setelah dilakukan pengujian, diperoleh akurasi tertinggi pada perbandingan data latih dan uji 70:30 sebesar 86,66%.

Kata kunci: *sifat kepribadian big five, klasifikasi, kepribadian, naïve bayes, twitter*

Abstract

In order to understand people's personality, their posts in social media can be used as a source of information. In this research, Naïve Bayes method is used to classify Twitter user's personality into one of the five classes, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Tweets were downloaded from 15 Twitter accounts using the Twitter API, with a total of 1,500 tweets. Each account is assigned a dominant personality trait based on the results of the personality questionnaire interpreted by a psychologist. Each tweet is pre-processed to lower case, followed by cleaning, tokenizing, stemming and removing stop words. Each word is weighted based on its frequency. Dataset is split into a training set and testing set with ratio 60:40, 70:30, 80:20 and 90:10. After testing, the highest accuracy of 86.66% was obtained on the ratio 70:30.

Keywords*: big five personality trait, classification, naïve bayes, personality, twitter*

1. Pendahuluan

Kepribadian adalah suatu pola watak yang relatif permanen dan sebuah karakter unik yang memberikan konsistensi sekaligus individualitas bagi perilaku seseorang [1]. Kepribadian dapat didefinisikan sebagai pola-pola perilaku, tata krama, pemikiran, motif dan emosi yang khas yang memberikan karakter kepada individu sepanjang waktu dan pada berbagai situasi yang berbeda [2]. Kepribadian dapat dinilai melalui pengukuran *self-report*, *peer-report* dan observasi pihak ketiga

Salah satu pendekatan yang dapat digunakan untuk memahami kepribadian adalah sifat (*trait*). Saat ini, banyak ahli psikologi yang berkeyakinan bahwa gambaran yang paling baik mengenai struktur sifat dimiliki oleh pendekatan kepribadian Big Five Personality Trait yang juga dikenal sebagai *five-factor model* atau OCEAN *model*. Model ini dikembangkan oleh Paul T. Costa, Jr. dan Robert R. McCrae. Sifat kepribadian yang ada pada model tersebut yaitu keterbukaan (*openness*), berhati-hati (*conscientiousness*), kenyamanan (*extraversi*), kebaikan (*agreeableness*), dan kestabilan emosi (*neuroticism*) [3]. Sifat-sifat kepribadian ini dapat dianalisa melalui kata-kata yang digunakan oleh seorang, yang tidak hanya dimengerti oleh para psikolog, namun dapat dimengerti oleh orang biasa [4]. Dengan demikian, kata-kata yang dituliskan oleh seorang pemilik akun media sosial dapat digunakan untuk memahami kepribadiannya.

Pengukuran *self-report* tidak praktis digunakan untuk menganalisis kepribadian di media sosial. Analisis kepribadian perlu dilakukan secara otomatis dengan menggunakan

pembelajaran mesin (*machine learning*), sehingga sifat kepribadian Big Five pemilik akun media sosial dapat diketahui tanpa perlu melakukan pengukuran *self-report* yang melibatkan pengisian kuesioner dan penilaian oleh pakar psikologi.

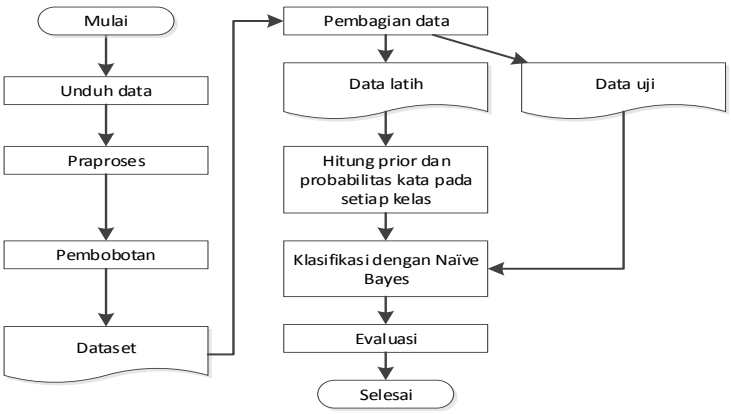
Qiu *et al.* [5] telah menunjukkan potensi pemanfaatan media sosial khususnya Twitter untuk penelitian kepribadian. Data dari Twitter digunakan oleh Sarwani dan Mahmudy [6] untuk mengklasifikasikan empat temperamen dasar (*artisan*, *guardian*, *idealist*, *rational*), serta digunakan oleh Ahmad dan Siddique [7] untuk mengklasifikasikan empat model kepribadian DISC (*dominance*, *influence*, *steadiness*, *compliance*).

Untuk sifat kepribadian Big Five dari pengguna Twitter, Kalghatgi *et al.* [8] menggunakan Multilayer Perceptron (MLP) Neural Network untuk mengklasifikasikan setiap sifat kepribadian, sementara Ong *et al.* [9] menggunakan Support Vector Machine dengan rataan akurasi tertinggi sebesar 76.23%. Pratama dan Sarno [10] membandingkan antara Naive Bayes, KNN and SVM, dimana Multinomial Naive Bayes memperoleh akurasi tertinggi dibandingkan metode-metode lainnya, sebesar 63%. Data dari Twitter juga digunakan oleh Golbeck *et al.* [11], Bai *et al.* [12], Quercia *et al.* [13], Damanik dan Khodra [14], Skowron *et al.* [15], Sharma dan Kaur [16], dan Guntuku *et al.* [17]. Penelitian-penelitian tersebut menggunakan analisis regresi untuk memprediksi nilai (skor) dari setiap sifat kepribadian Big Five pengguna Twitter.

Dalam penelitian ini, dilakukan klasifikasi kepribadian pengguna Twitter ke dalam lima kelas, yaitu Openness, Conscientiousness, Extraversion, Agreeableness, dan Neuroticism. Klasifikasi dilakukan untuk menemukan sifat kepribadian yang dominan di antara sifat-sifat lainnya. Metode klasifikasi yang digunakan adalah Naïve Bayes.

2. Metode Penelitian

Tahapan-tahapan yang dilakukan dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan penelitian

Data yang digunakan dalam penelitian ini adalah *tweet* dari akun Twitter mahasiswa-mahasiswa di Jurusan Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau. Setiap akun ditetapkan kepribadiannya berdasarkan kuesioner yang telah divalidasi oleh seorang psikolog. Kuesioner diadaptasi dari standar baku IPIP (International Personality Item Pool - Five Factor Inventory). Psikolog menginterpretasikan kepribadian seseorang sebagai sifat yang paling dominan di antara kelima sifat kepribadian yang ada.

Tweet diperoleh dengan cara mengunduhnya melalui Twitter API. *Tweet* yang dikumpulkan tidak termasuk *retweet* yang dibuat oleh akun lain. Setiap *tweet* dipraproses menjadi huruf kecil (*case folding*), dihilangkan karakter-karakter yang tidak penting (*cleaning*) termasuk menghilangkan entitas *tweet* (*mention*, *hashtag*, URL) dan *emoticon*, ditokenisasi menjadi kata (*word tokenization*), kemudian ditemukan kata dasarnya (*stemming*). *Stemming* dilakukan dengan menggunakan algoritma Enhanced Confix Stripping (ECS). Selanjutnya, dihilangkan kata-kata yang tidak penting (*stopword removal*). Setiap kata dibobot berdasarkan Term Frequency (TF).

Sebelum dilakukan klasifikasi, *dataset* dibagi menjadi kumpulan data latih (*training set*) dan kumpulan data uji (*testing set*) dengan perbandingan 60:40, 70:30, 80:20 dan 90:10.

Klasifikasi dilakukan dengan menggunakan metode Naïve Bayes. Akhirnya, dilakukan pengujian akurasi dengan menggunakan *confusion matrix*.

3. Hasil dan Diskusi

Responden kuesioner meliputi 95 orang mahasiswa jurusan Teknik Informatika. Setelah diisi oleh mahasiswa, kuesioner-kuesioner tersebut diserahkan kembali kepada psikolog untuk dilakukan analisa dan dinilai jawaban dari setiap pertanyaan. Hasil kuesioner kepribadiannya diperlihatkan pada Tabel 1.

Tabel 1. Hasil kuesioner

Kepribadian	Jumlah (orang)
Openness	30
Conscientiousness	31
Extraversion	6
Agreableness	25
Neuroticism	3
Total	95

Berdasarkan Tabel 1, diketahui jumlah orang per kepribadian tidak seimbang. Untuk memperoleh jumlah yang seimbang, diambil data per kepribadian sebanyak nilai terendah, yaitu 3. Masing-masing kepribadian diambil sebanyak 3 orang, sehingga total akun yang akan diunduh sebanyak 15 akun Twitter. *Tweet* diunduh dengan menggunakan Twitter API, sebanyak 100 *tweet* per akun dengan total keseluruhan sebanyak 1.500 *tweet*.

Setiap *tweet* diberi label kelas sesuai kepribadian si pemilik akunnya. Selanjutnya, teks pada *tweet* dipraproses dan kata-katanya dibobot. Klasifikasi dengan menggunakan Naïve Bayes dilakukan pada perbandingan jumlah data latih dan uji sebesar 60:40, 70:30, 80:20, 90:10. Pembagian data dilakukan secara acak dengan jumlah data per label kelas diupayakan berimbang. Hasil pengujian berupa label kelas dari setiap data uji. Sifat kepribadian yang paling dominan diambil dari label kelas dengan frekuensi terbesar. Dengan kata lain, Akun X dinyatakan memiliki kepribadian Y karena jumlah *tweet*-nya yang berlabel Y lebih banyak dibandingkan keempat label lainnya.

Pada setiap pembagian data, pengujian dilakukan empat kali dengan menggunakan *dataset* yang telah dimodifikasi pada tahapan praprosesnya. Berdasarkan langkah *stemming*, ada dua kombinasi. Pertama, jika suatu kata tidak ditemukan kata dasarnya, maka kata tersebut tidak dikembalikan sehingga kata tersebut tidak digunakan pada langkah selanjutnya. Kedua, kata tersebut dikembalikan apa adanya. Berdasarkan langkah *stopword removal*, ada dua kombinasi. Pertama, tidak melakukan penghilangan kata-kata yang tidak penting. Kedua, dilakukan penghilangan.

Hasil pengujian dimana kata yang tidak memiliki kata dasar tidak digunakan, serta tidak dilakukan penghilangan kata tidak penting diperlihatkan pada Tabel 2.

Tabel 2. Pengujian pertama

Perbandingan Latih:Uji	Akurasi
60:40	60,00%
70:30	86,66%
80:20	60,00%
90:10	46,66%

Hasil pengujian dimana kata yang tidak memiliki kata dasar digunakan apa adanya, serta tidak dilakukan penghilangan kata tidak penting diperlihatkan pada Tabel 3.

Tabel 3. Pengujian kedua

Perbandingan Latih:Uji	Akurasi
60:40	66,66%
70:30	73,33%
80:20	60,00%
90:10	73,66%

Hasil pengujian dimana kata yang tidak memiliki kata dasar tidak digunakan, serta dilakukan penghilangan kata tidak penting diperlihatkan pada Tabel 4.

ISSN (Printed) : 2579-7271
ISSN (Online) : 2579-5406

Tabel 4. Pengujian ketiga

Perbandingan Latih:Uji	Akurasi
60:40	60,00%
70:30	66,66%
80:20	46,66%
90:10	60,00%

Hasil pengujian dimana kata yang tidak memiliki kata dasar digunakan apa adanya, serta dilakukan penghilangan kata tidak penting diperlihatkan pada Tabel 5.

Tabel 5. Pengujian keempat

Perbandingan Latih:Uji	Akurasi
60:40	66,66%
70:30	73,66%
80:20	80,00%
90:10	53,33%

Selain itu, dilakukan prediksi terhadap tiga akun Twitter di luar *dataset*. Berdasarkan hasil prediksi, diperoleh akurasi sebesar 66,66%, dimana terdapat dua akun yang sesuai dengan hasil kuesioner kepribadian, sedangkan satu akun tidak sesuai.

4. Kesimpulan

Dalam penelitian ini, diperlihatkan bahwa metode Naïve Bayes dapat digunakan untuk mengklasifikasikan kepribadian Big Five dari pengguna Twitter berdasarkan *tweet*-nya. Berdasarkan hasil pengujian, diperoleh akurasi tertinggi pada perbandingan data latih dan uji 70:30 sebesar 86,66%. Hasil pengujian juga memperlihatkan pengaruh tahapan praproses terhadap akurasi, dimana kata yang tidak memiliki kata dasar tidak digunakan, serta tidak dilakukan penghilangan kata tidak penting

Kedepannya, penelitian ini tidak hanya menggunakan kata-kata yang ditemukan kata dasarnya. Perlu dilakukan normalisasi terhadap kata-kata yang ditulis tidak sesuai aturan berbahasa, baik sengaja maupun tidak sengaja, untuk mengetahui dampaknya terhadap hasil akurasi. Selain itu, dapat menggunakan fitur lain sebagaimana tercantum pada [18] dan [19]. Dengan semakin banyaknya jumlah fitur yang akan digunakan, disarankan untuk menerapkan langkah seleksi ciri dengan menggunakan F-score. Untuk metode klasifikasi yang digunakan, disarankan untuk membandingkan performa Naïve Bayes terhadap k-Nearest Neighbor (k-NN) dan Support Vector Machine (SVM) pada *dataset* yang sama.

Referensi

[1] Feist J, Feist GJ. Theories of Personality. 6 ed. Yogyakarta: Pustaka Pelajar. 2008.
[2] Wade C, Tavris C. Psikologi. 9 ed. Jakarta: Erlangga. 2007.
[3] Santrock JW. Psikologi Pendidikan (Educational Psychology). Jakarta: Salemba Humanika 2011.
[4] Febrianto N, Prasetia I, Wijaya A. Pembuatan Sistem Prediksi Kepribadian “The Big Five Traits” dari Media Sosial Twitter. Jakarta: Universitas Bina Nusantara; 2015.
[5] Qiu L, Han L, Ramsay J, Yang F. You are what you tweet: Personality expression and perception on Twitter. Journal of Research in Personality 2012;46:710-8.
[6] Sarwani MZ, Mahmudy WF. Analisis Twitter Untuk Mengetahui Karakter Seseorang Menggunakan Algoritma Naïve Bayes Classifier. Seminar Nasional Sistem Informasi Indonesia; 2-3 November 2015; Surabaya2015.
[7] Ahmad N, Siddique J. Personality Assessment using Twitter Tweets. 21st International Conference on Knowledge Based and Intelligent Information and Engineering Systems; Marseille, France: Elsevier, B.V.; 2017.
[8] Kalghatgi MP, Ramannavar M, Sidnal NS. A Neural Network Approach to Personality Prediction based on the Big-Five Model. International Journal of Innovative Research in Advanced Engineering. 2015;8(2):56-63.
[9] Ong V, Rahmanto ADS, Williem, Suhartono D, Nugroho AE, Andangsari EW, et al., editors. Personality Prediction Based on Twitter Information in Bahasa Indonesia. Federated Conference on Computer Science and Information Systems (FedCSIS); 2017 3-6 November 2017; Prague, Czech Republic: IEEE.
[10] Pratama BY, Sarno R, editors. Personality Classification Based on Twitter Text Using Naive Bayes, KNN and SVM. International Conference on Data and Software Engineering; 2015 25-26 November 2015; Yogyakarta, Indonesia: IEEE.

[11] Golbeck J, Robles C, Edmondson M, Turner K. Predicting Personality from Twitter. IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing2011.
[12] Bai S, Yuan S, Hao B, Zhu T, editors. Predicting Personality Traits of Microblog Users. IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT); 2013 17-20 November 2013; Atlanta, GA, USA: IEEE.
[13] Quercia D, Kosinski M, Stillwell D, Crowcroft J, editors. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing; 2011 9-11 October 2011; Boston, MA, USA: IEEE.
[14] Damanik AT, Khodra ML. Prediksi Kepribadian Big 5 Pengguna Twitter dengan Support Vector Regression. Jurnal Cybermatika. 2015;3(1):14-22.
[15] Skowron M, Ferwerda B, Tkalcic M, Schedl M. Fusing Social Media Cues: Personality Prediction from Twitter and Instagram. the 25th International Conference Companion on World Wide Web. 2016:107-8.
[16] Sharma K, Kaur A. Personality prediction of Twitter users with Logistic Regression Classifier learned using Stochastic Gradient Descent. IOSR Journal of Computer Engineering. 2015;17(4):39-47.
[17] Guntuku SC, Lin W, Carpenter J, editors. Studying Personality through the Content of Posted and Liked Images on Twitter. the 2017 ACM on Web Science Conference; 2017 25-28 Juni 2017; Troy, New York, USA: ACM.
[18] Riquelme F, Gonzalez-Cantergiani P. Measuring User Influence on Twitter - a Survey. Journal of Information Processing and Management. 2016.
[19] Pal A, Counts S. Identifying Topical Authorities in Microblogs. WSDM'11; 9-12 February 2011; Hong Kong, China: ACM; 2011.