Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 988-996
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

Prediksi Kepribadian Berdasarkan Media Sosial *Twitter* Menggunakan Metode *Naïve Bayes Classifier*

Muhammad Ichsanudin A[1], Agung Susilo Yuda Irawan[2], Arip Solehudin[3]
[1,2,3]Program Studi Teknik Informatika, Universitas Singaperbangsa Karawang
Jl. HS. Ronggo Waluyo Teluk Jambe Timur Karawang 41361
muhammad.ichsanudin17150@student.unsika.ac.id, agung@unsika.ac.id,
arip.solehudin@staff.unsika.ac.id

Abstract

Analysis of a person's personality is very helpful as an assessment in various matters such as recruitment, career, health. The methods commonly used in personality analysis are interviews, observations, and questionnaire surveys. This study tries to provide a solution by simply using social media, namely twitter, by analyzing twitter user information data called tweets, this is to add to the method of personality analysis. The method used in this personality prediction research is to classify a tweet into 5 personality forms. The personality method used by the researcher is the Big Five Personality which consists of openness, conscientiousness, extraversion, agreeableness, and neuroticism with classification calculations using Naive Bayes. The result of this research is an accuracy of 42% with the highest class, namely Agreeableness.

Keywords: *Data mining, Text mining, Twitter, Big Five, Naive Bayes*

Abstrak

Analisis kepribadian seseorang sangat membantu sebagai penilaian dalam berbagai hal seperti perekrutan, karir, kesehatan. Metode yang biasa digunakan dalam analisis kepribadian dengan cara wawancara, observasi, dan survei kuesioner. Penelitian ini mencoba memberi solusi dengan cukup menggunakan media sosial yaitu twitter, dengan menganalisa data informasi pengguna twitter yang disebut tweet, hal ini untuk menambah metode dari analisis kepribadian. Metode yang digunakan dalam penelitian prediksi kepribadian ini dilakukan untuk mengklasifikasi sebuah tweet kedalam bentuk 5 kepribadian. Metode kepribadian yang digunakan peneliti adalah Big Five Personality yang terdiri dari openness, conscientiousness, extraversion, agreeableness, dan neuroticism dengan perhitungan klasifikasi dengan Naive Bayes. Hasil dari penelitian ini adalah akurasi sebesar 42% dengan kelas terbanyak yaitu Agreeableness.

Kata kunci: *Data mining, Text mining, Twitter, Big Five, Naive Bayes*

1. PENDAHULUAN

Setiap individu mempunyai satu kesatuan yang utuh dan unik. Hal ini menandakan bahwa kesatuan utuh dalam setiap individu tidak dapat dipisahkan dengan segala cirinya, oleh karena itu individu mempunyai kepribadian yang berbeda –beda. Dalam kepribadian terdapat banyak teori yang telah dikemukakan oleh beberapa ahli salah satu nya adalah teori big five personality yang mencakup lima kepribadian besar. *Big five Personality* atau model lima besar sifat kepribadian merupakan salah satu teori yang dikemukakan oleh Lewis Goldberg.Menurut lewis ada lima model yang bisa mendifinisikan kepribadian manusia adapun lima kepribadian tersebut yaitu *Openness to Experience*, *Conscientiousness*, *Extraversion*, *Agreeableness*, dan

Neuroticism. Orang *openness* memiliki ciri positif ialah cenderung lebih kreatif, imajinatif, intelektual, penasaran serta berpikiran luas. Orang *Conscientiousness* memiliki ciri positif ialah bisa diandalkan,bertanggung jawab, tekun serta berorientasi pada pencapain. Ciri Positif Orang *Extraversion* merupakan bahagia berteman, gampang bersosialisasi,hidup berkelompok serta tegas. Ciri positif *Agreableness* merupakan kooperatif (bisa berkolaborasi), penuh keyakinan, bertabiat baik, hangat serta berhati lembut dan suka menolong. Ciri positif orang *Neuroticism* disebut dengan *Emotional Stability* (Stabilitas Emosional), Individu dengan Emosional yang stabil cenderang tenang saat menghadapi masalah, percaya diri, memiliki pendirian yang teguh. (Faidh, Erwin, 2019).

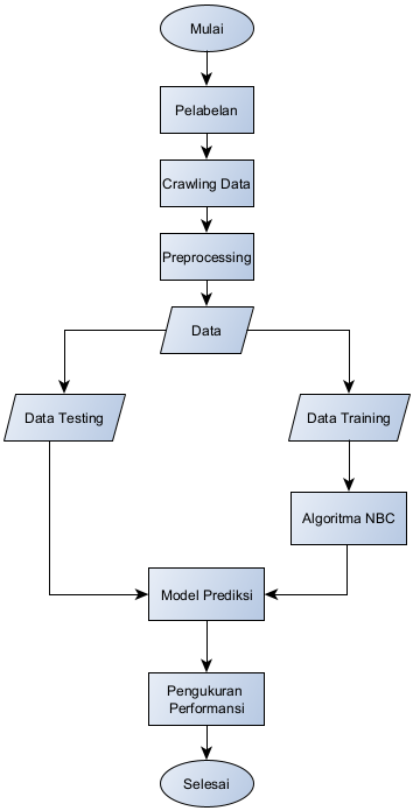
Terdapat banyak cara untuk menentukan kepribadian seseorang salah satunya melalui media sosial *Twitter*. Twitter merupakan suatu web jejaring sosial yang lagi tumbuh pesat dikala ini sebab pengguna bisa berhubungan dengan pengguna yang lain dari pc maupun fitur mobile mereka dari manapun serta kapanpun. *Twitter* selaku suatu web jejaring sosial membagikan akses kepada penggunanya buat mengirimkan suatu pesan pendek yang terdiri dari maksimal 140 kepribadian(diucap *tweet*).

Universitas Singaperbangsa Karawang adalah satu-satunya universitas negeri yang berada di kota Karawang. Universitas Singaperbangsa Karawang terdiri dari beberapa fakultas seperti fakultas teknik, fakultas agama islam, fakultas ilmu pendidikan, fakultas ilmu komputer, fakultas ilmu kesehatan fakultas hukum dan lain sebagainya. Dari berbagai fakultas tersebut mengkaji ilmu-ilmu yang sangat bermanfaat contohnya fakultas ilmu komputer yang mengkaji ilmu komputer. Oleh karena itu peneliti ingin menerapkanya pada mahasiswa Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang tentunya dengan harapan akan berguna di masa yang akan datang. Suatu metode yang mampu mengklasifikasi kepribadian seseorang dengan tingkat akurasi yang cukup tinggi, salah satunya adalah metode *Naive Baiyes* dari penelitian yang berjudul "Klasifikasi Kepribadian Big Five Pengguna Twitter dengan Metode Naïve Bayes". Penelitian ini merupakan pengembangan dari penelitian tersebut dengan menggunakan metode *Naive Bayes Classifier*. Dengan harapan mampu mengklasifikasikan kepribadian dari data yang sangat banyak menjadi lebih efektif dan efisien yang tentunya dengan tingkat akurasi yang tinggi

2. METODOLOGI PENELITIAN

Metode yang digunakan dalam penelitian ini adalah Algoritma *Naive Bayes Classifier* dengan beberapa tahapan yaitu, Pelabelan *dan Crawling Data, Data Preprocessing,* Klasifikasi *Naive Bayes*, Pengukuran Performansi.

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 988-996
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti



Gambar 1. Rancangan Sistem

2.1. Pelabelan

Pada proses ini dicoba pemberian label kelas jadi lima kelas karakter *big five* dari hasil kuesioner yang sudah disebar ke koresponden dari 30 persoalan bersumber pada penilai BFI (*Big Five Inventory*). Hal ini dilakukan untuk mengetahui jenis label kepbradiannya dan nama *Twitter* nya.

2.2. *Crawling Data*

Crawling Data ialah sesi yang bertujuan buat memperoleh informasi yang hendak digunakan sebagai informasi acuan oleh sistem yang berbentuk user serta tweet.

2.3. *Data Preprocessing*

Proses pada sesi preprocessing informasi ini bertujuan buat memperoleh informasi acuan yang siap diproses ke dalam sistem klasifikasi dari informasi tweet mentah yang diganti ke dalam wujud yang lebih simpel. Preprocessing informasi yang dicoba terhadap informasi tweet merupakan selaku berikut, *Case Folding, Tokenizing, Filtering, Stemming.*

2.4. Klasifikasi *Naive Bayess*

Proses ini ialah proses utama yang bertujuan buat mengklasifikasikan informasi yang telah melewati proses tadinya dengan memakai tata cara Naïve Bayes Classifier.

2.5. Model Prediksi

Proses ini ialah sistem pendidikan yang telah terbuat buat menciptakan model prediksi karakter *big five*.

2.6. Pengukuran Performansi

Proses ini ialah sesi akhir ialah menghitung tingkatan akrurasi dari sistem yang telah terbuat untuk menghitung akurasi dari hasil algoritma *Naive Bayess Classifier*

3. HASIL DAN PEMBAHASAN

Pada bagian ini dijelaskan hasil uji dan analisis dari sistem yang telah dibangun sesuai dengan metodelogi penelitian.

3.1. Pelabelan

Dataset yang digunakan dalam penelitian ini berasal dari hasil penyebaran kuesioner peneliti terhadap beberapa mahasiswa yang meliputi nama twitter dan berisi 30 pertanyaan berdasarkan standar internasional IPIP (*International Personality Item Pool*) dengan total sebanyak 45 mahasiswa yang mengisi kuesioner tersebut. Kemudian, data hasil pengisian kuesioner tersebut diseleksi untuk mencari mahasiswa yang mempunyai *twitter* dan akun twitter nya tidak diprivate dari total 45 data menjadi 18 data yang siap digali tweetnya. Berikut contoh data yang sudah diseleksi.

Tabel 1. Data Kuesioner

No.	From-User	Label
1	iyaaaa	Agreeableness
2	Ayu Shafira Tubagus	Openness to Experience
3	Daffaaaa	Extraversion
4	dinmarf	Conscientiousness
5	dyas	Conscientiousness
6	Endah Sulistyo N	Agreeableness
7	Fachry Ikhsal	Openness to Experience
8	dandelions	Agreeableness
...	....	....
18	nci	Agreeableness

3.2. Crawling Data

Pada proses ini dilakukan crawling data tweet dari data hasil seleksi kuesioner dengan masing-masing pengguna maksimal 100 tweet. Pada proses crawling ini menggunakan aplikasi rapidminer Studio dan memerlukan nama pengguna twitter dan pengguna twitter yang akunya tidak di private. Peneliti menseleksi data dan hanya mengambil 2 atribut yaitu From-User dan Text. Berikut hasil data crawling twitter yang sudah di sematkan ke excel.

Tabel 2. *Data Crawling Twitter*

No.	From-User	Text
1	iyaaaa	#NewProfilePic https://t.co/vkTogt3DWU
2	iyaaaa	RT @bertanyarl: Tanyarl mohon bantuannya ya teman-teman ?? https://t.co/9iCIfgngHI
3	iyaaaa	@amaliaprnmsr_ @txtdarigajelas @anisanurs22 @mupey_vale Wkwkwk skip moal dibotak ceuk aku ge
4	iyaaaa	RT @infoBMKG: Peringatan dini cuaca wilayah Kalimantan Timur [14 April 2021] #BMKG Selengkapnya klik tautan berikut https://t.co/w0wq8gPZFo
5	iyaaaa	Terimakasih kepada @bpptik @kemkominfo yang telah menyelenggarakan sertifikasi pelatihan secara gratis, dengan tenaga pengajar yang baik, materi yang sangat baik dan sangat berguna untuk kedepanya banyak ilmu yang didapat dari #bpptik
6	iyaaaa	Anggap saja sedang curhat, tapi ada yang nemenin bedanya.
7	iyaaaa	Tapi sebenernya yang lu rasain ketika ngeliat cuitan orang lain dan 'dih anying sama' nahh mulai cocokologi dengan si cuit, dan merasa paling emm dan lupa bahwa pemeran lain juga pernah emm buat lu, thats why aing kadang males buka twit
...	...	.....
1674	Nanu banget	@karawangfess Ngapainn maluuuuu, kalo malu mah lu belum kenal udah jadian
1675	Nanu banget	Haaaa

Data diatas memiliki 2 atribut yaitu From-user dan Text dengan total jumlah 1675 tweet dari masing-masing akun maksimal tweet yang diambil adalah 100 tweet. Tweet tersebut dimasukan kedalam attribut text yang nantinya akan dilanjutkan ke tahap preprocessing.

3.3. *Data Preprocessing*

Pada proses ini adalah tahapan agar data teks dapat diubah menjadi lebih terstruktur peneliti melakukan beberapa tahapan untuk memproses data teks yang terdiri dari *Tokenizing, Case folding, Filtering, dan Stemming.*

1. ***Tokenizing***
   Pada tahapan ini dilakukan pemotongan string input berdasarkan tiap kata yang menyusunnya. Namun untuk karakter petik tunggal (‘), titik (.), semikolon (;), titk dua (:) atau lainnya, dapat memiliki peran yang cukup banyak sebagai pemisah kata. Berikut merupakan tampilan data yang sudah melewati proses *tokenizing*.
2. ***Case Folding***
   Pada tahapan ini dilakukan proses pengubahan huruf kapital menjadi huruf kecil *(lower case).* Sebagai contoh, *user* yang ingin mendapatkan informasi “KOMPUTER” dan mengetik “KOMPOTER”, “KomPUter” atau “komputer”, tetap diberikan hasil retrieval yang sama yakni “komputer”. Hanya huruf ‘a’ sampai dengan ‘z’ yang diterima. Karakter selain huruf dihilangkan dan dianggap delimiter.
3. ***Filtering***
   Pada tahapan ini mengambil kata-kata penting dari hasil token. algoritma stoplist (membuang kata kurang penting). atau wordlist (menyimpan kata penting). Stoplist/stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah “yang”, “dan”, “di”, “dari” dan seterusnya.
4. ***Stemming***
   Pada tahapan ini dilakukan untuk memperkecil jumlah indeks yang berbeda dari suatu dokumen, juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk atau *form* yang berbeda karena mendapatkan imbuhan yang berbeda. Sebagai contoh kata “dasarnya” , “berdasarkan”, “ mendasari” akan distem ke *root word*-nya yaitu “dasar”.

Tahapan preprocessing ini perlu dilakukan untuk menseleksi data agar data menjadi lebih tersturktur lagi yang nantinya akan diproses ke tahapan selanjutnya. Berikut tampilan contoh data yang sudah melalui tahap preprocessing.

Tabel 3. Data Hasil *Text Processing*

From-User	Label	aku	banget	bicara	dapat	dasar	....	zzYUGxmZe
iyaaaa	Agreeableness	00.00	00.00	00.00	0.348735	00.00	....	00.00
iyaaaa	Agreeableness	00.00	00.00	00.00	00.00	00.00	....	00.00
iyaaaa	Agreeableness	00.00	00.00	00.00	00.00	00.00	....	00.00
iyaaaa	Agreeableness	00.00	00.00	00.00	00.00	00.00	....	00.00

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 988-996
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti

From-User	Label	aku	banget	bicara	dapat	dasar	....	zzYUGxmZe
iyaaaa	Agreeablenes s	00.00	00.00	00.00	00.00	00.00	....	00.00
.....	.....	....	...	...	...	..	...	...
Nanu banget	Neuroticism	00.00	00.00	00.00	00.00	00.00	....	00.00
Nanu banget	Neuroticism	00.00	00.00	00.00	00.00	00.00	....	00.00

Pada tabel diatas menunjukan tiap tweet berapa kali jumlah kata yang keluar dari tiap attribut kata yang nantinya akan dijadikan informasi.

3.4. Klasifikasi *Naive Bayes*

Pada tahap ini, algoritma yang digunakan adalah *Naive Bayes* dengan memanfaatkan perhitungan probabilitas dan statistik berdasarkan probabilitas sebelumnya. Pada tahap ini peneliti memisah data untuk dijadikan data training dan data testing dengan perbandingan 70:30. Untuk menghitung probabilitas menggunakan *rapidminer Studio*, sehingga hasil yang didapat dapat dilihat dari gambar berikut.

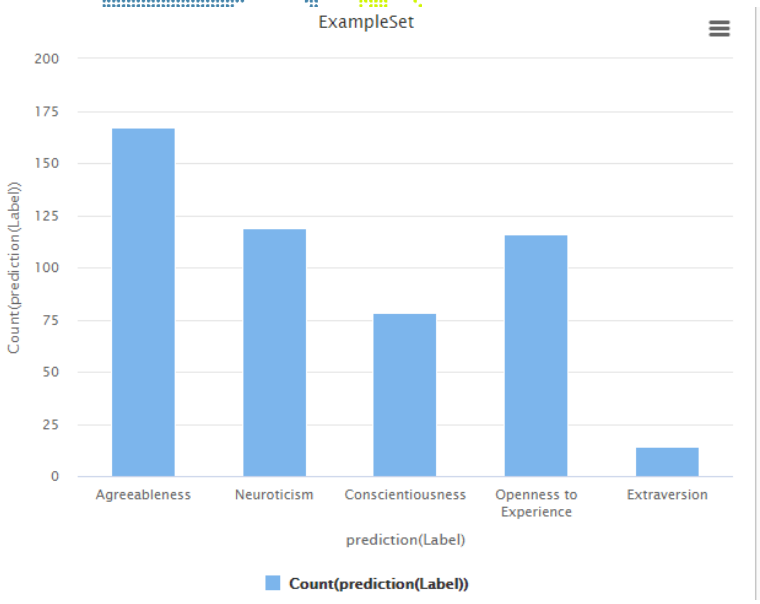
SimpleDistribution

```
Distribution model for label attribute Label


Class Agreeableness (0.345)
5098 distributions

Class Openness to Experience (0.390)
5098 distributions

Class Extraversion (0.068)
5098 distributions

Class Conscientiousness (0.160)
5098 distributions

Class Neuroticism (0.037)
5098 distributions
```

Gambar 2. Hasil *NBC*

Dari gambar diatas diketahui penyebaran untuk attribut label dari 5098 *attribut* kata dan terbagi kedalam 5 kelas. Untuk lebih jelasnya bisa dilihat dari grafik berikut.

Jurnal Sains Komputer & Informatika (J-SAKTI)
Volume 5 Nomor 2, September 2021, pp. 988-996
ISSN: 2548-9771/EISSN: 2549-7200
https://tunasbangsa.ac.id/ejurnal/index.php/jsakti



Gambar 3. Grafik Penyebaran Label

Dapat dilihat dari gambar 3 diatas kelas tertinggi dari hasil prediksi adalah kelas *Agreeableness* dengan jumlah 167 dan kelas terendah adalah kelas *Extraversion* dengan jumlah 14.

3.5. Pengukuran Performansi

Setelah dilakukanya perhitungan dengan algoritma *Naive Bayes* tahap selanjutnya dilakukan uji akurasi. Hal ini dilakukan untuk mengukur seberapa akurat algoritma *Naive Bayes* dalam melakukan prediksi. Dalam tahap ini yang di uji adalah akurasi, *recall*, dan *precision*. Untuk tampilannya bisa dilihat pada gambar tabel berikut.

PerformanceVector

```
PerformanceVector:
accuracy: 42.71%
ConfusionMatrix:
True:   Agreeableness   Openness to Experience  Extraversion    Conscientiousness       Neuroticism
Agreeableness:  105     32      10      15      5
Openness to Experience: 21      61      11      17      6
Extraversion:   2       5       6       1       0
Conscientiousness:      21      20      2       33      2
Neuroticism:    21      75      4       13      6
weighted_mean_recall: 36.98%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   Agreeableness   Openness to Experience  Extraversion    Conscientiousness       Neuroticism
Agreeableness:  105     32      10      15      5
Openness to Experience: 21      61      11      17      6
Extraversion:   2       5       6       1       0
Conscientiousness:      21      20      2       33      2
Neuroticism:    21      75      4       13      6
weighted_mean_precision: 41.13%, weights: 1, 1, 1, 1, 1
ConfusionMatrix:
True:   Agreeableness   Openness to Experience  Extraversion    Conscientiousness       Neuroticism
Agreeableness:  105     32      10      15      5
Openness to Experience: 21      61      11      17      6
Extraversion:   2       5       6       1       0
Conscientiousness:      21      20      2       33      2
Neuroticism:    21      75      4       13      6
```

Gambar 4. Hasil Pengukuran Performansi

Pada gambar diatas diketahui akurasi yang didapat adalah 42,71% diketahui jumlah *recall* yang didapat adalah 36,98% dan jumlah *precision* yang didapat adalah 41,13%.

4. SIMPULAN

Dari hasil penelitian yang dilakukan ini dapat disimpulkan penelitian kepribadian pengguna Twitter dapat diprediksi melalui tweet pengguna twitter walaupun dengan jumlah akurasi yang masih tergolong rendah sebesar 42%. Berdasarkan hasil prediksi *label* kepribadian tertinggi ada pada *label Agreeableness* dengan jumlah 167. Sedangkan, nilai terendah ada pada *label Extraversion* dengan jumlah 14 *tweet*. Hal ini dikarenakan pada *data training* dengan jumlah 1173 data *tweet* lebih dominan berlabel *Agreeableness* maka dari itu model yang dibangun cenderung memprediksi *label Agreeableness* dan membuat nilai *recall* dan *precision* pada kelas *Agreeableness* meningkat tetapi tidak pada kelas yang lainnya.

Saran untuk penelitian selanjutnya, dalam penelitian ini jumlah tweet maksimal per akun berjumlah 100 tweet dan ada beberapa akun yang jumlah tweet nya tidak mencapai 100. Maka dari itu untuk penelitian selanjutnya diharapkan data tweet per akun jumlahnya harus sama rata agar mendapat nilai yang lebih maksimal. untuk penelitian selanjutnya menggunakan data yang seimbang antar kelas agar untuk membuat model yang dibangun tidak cenderung memprediksi jenis kelas yang terbanyak.

DAFTAR PUSTAKA

[1] Karthika R, R. Kathiyayini. “Facebook Data for Predictive Personality Analysis”. IJSRCSEIT., vol.2, no.2, pp. 1147-1150. 2017.

[2] Y B N D Artissa, I Asror, S A Faraby. “Personality Classification based on Facebook status text using Multinomial Naïve Bayes method”. Journal of Physics: Conference Series., vol.2, pp. 1-8. 2019.

[3] Haq, F. I. N., & Setiawan, E. B. (2019). “Implementasi Naive Bayes Classifier untuk Prediksi Kepribadian Big Five pada Twitter Menggunakan Term Frequency-Inverse Document Frequency (TF-IDF) dan Term Frequency-Relevance Frequency (TF-RF)”. eProceedings of Engineering, 6(2).

[4] Ellandi, R., Setiawan, E. B., & Nugraha, F. N. (2019).” Prediksi kepribadian Big Five dengan Term-Frequency Inverse Document Frequency Menggunakan Metode k-Nearest Neighbor pada Twitter”. eProceedings of Engineering, 6(2).

[5] Yusra, Y., Fikry, M., Syarfianto, R., Candra, R. M., & Budianita, E. (2018, November). “Klasifikasi Kepribadian Big Five Pengguna Twitter dengan Metode Naïve Bayes”. In Seminar Nasional Teknologi Informasi Komunikasi dan Industri (pp. 317-321).

[6] Han, J. dan Kamber, M., 2006, “Data Mining: Concepts and Techniques-Chapter 2”. USA: Elsevier