# Prediksi kepribadian *Big Five* dengan *Term-Frequency Inverse Document Frequency* Menggunakan Metode *k-Nearest Neighbor* pada *Twitter*

Roji Ellandi[1], Erwin Budi Setiawan S.Si., M.T[2], Dr. Fida Nirmala Nugraha, s.Psi.,M.Psi.[3]

[1,2,3]**Fakultas Informatika, Universitas Telkom, Bandung**
[1]rojiellandi@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id,
[3]fidanurmalanugraha@telkomuniversity.ac.id

**Abstrak**

***Twitter*** **merupakan media sosial yang sampai saat ini sangat digemari dan menjadi penyebaran informasi yang sangat cepat. Informasi yang beredar juga sangat banyak, mulai dari berita, opini, komentar, dan kritik semuanya ada yang bersifat positif , negatif, dan netral. Menurut data yang dilansir secara statistik dan berdasarkan penelitian *PeerReach*, indonesia termasuk pengguna *Twitter* yang paling aktif ke 3 di dunia dibawah Amerika serikat dan Jepang. Dari kumpulan data tersebut kita dapat melakukan analisis kepribadian terhadap suatu keadaan untuk melihat respon masyarakat, media ataupun kepemerintahan terhadap suatu objek tersebut dan proses klasifikasi itu sendiri. Metode yang digunakan dalam penelitian prediksi kepribadian ini dilakukan untuk mengklasifikasi sebuah *tweet* ke dalam bentuk 5 kepribadian. Metode kepribadian yang digunakan peneliti adalah *Big Five Personality* dan metode perhitungan klasifikasi dengan k-NN (*k-Nearest Neighbor*). hasil dari penelitian ini adalah dapat memperoleh nilai akurasi 60,97% dengan pembobotan melalui tahap TF-IDF(*Term-Frequency Invert Document Frequency*).**

**Kata kunci : Klasifikasi, *Big Five Personality*, *Twitter*, *TF-IDF*, *k-NN***

**Abstract**

**Twitter is a social media that has been very popular and has become a very fast dissemination of information. There is also a lot of information circulating, ranging from news, opinions, comments, and criticism all of which are positive, negative, and neutral. According to data reported statistically and based on PeerReach's research, Indonesia is the 3rd most active Twitter user in the world under the United States and Japan. From the data collection we can conduct personality analysis of a situation to see the response of the community, the media or governance of an object and the classification process itself. The method used in personality prediction research is done to classify a tweet into 5 personality forms. The personality method used by researchers is the Big Five Personality and the k-NN (k-Nearest Neighbor) classification method. the results of this study were able to obtain an accuracy value of 60,97% by weighting through the TF-IDF (Term-Frequency Invert Document Frequency) stage**

**Keyword : Classification, Big Five Personality, Twitter, TF-IDF, k-NN**

## 1.Pendahuluan

Kepribadian Big five merupakan salah satu metode yang dikenal dalam dunia pisikologi untuk menginterpretasikan kepribadian seseorang, terutama untuk menemukan hubungan kepribadian dengan lingkungan pekerjaan. Kepribadian *big five* terdiri dari *Openness to Experience(O), Conscientiosness(C), Extraversion(E), Agree-ableness(A),* dan *Neuroticism(N)* (Costa P. Dkk, 1991), keperibadian O memiliki imajinasi yang aktif,kepekaan terhadap estetika, kepedulian terhadap perasaan pribadi, ketertarikan terhadap perbedaan, keingintahuan intelektual, dan kebebasan berpendapat, kepribadian C berhubungan erat dengan mengendalikan *impulse*, pengendalian diri demi perencanaan yang matang, pengaturan, dan pengerjaan tugas tugas. Kepribadian E percaya diri, aktif, cerewet, optimis, serta menyukai kesenangan dan selalu merasakan ceria secara alami. Kepribadian A mengutamakan orang lain, simpatik terhadap orang lain, dan suka menolong. Kepribadian N cenderung mengalami perasaan negatif seperti ketakutan, kesedihan, rasa canggung, kemarahan, serta rasa bersalah dan suka benci.

Kepribadian berhubungan dan mempengaruhi beberapa aspek dari linguistik. Prediksi berdasarkan linguistik dilakukan dengan menganalisis pemilihan kata kata dan letak kata tersebut di dalam katagori yang di tentukan sesuai dengan bahasa yang digunakan. Analisis linguistik telah dilakukan terhadap beberapa Profil media sosial dan penggunaan bahasa sehari hari, pesan singkat. Para psikologi dengan menemukan korelasi berbagai variabel linguistik dengan

kepribadian. Beberapa perusahaan khususnya perusahaan industri menengah ke atas telah menggunakan media sosial untuk mempertimbangkan penerimaan pegawai baru, selain hasil test psikologi formal yang selalu dilakukan (*CareerBuilder*,2012). Berdasarkan hal ini, kepribadian seseorang dapat diprediksi berdasarkan informasi seseorang melewati akun media sosial, seperti *facebook* atau *Twitter*. Pada penelitian ini bertujuan membangun model prediksi kepribadian *Big five personality* dari penggunaan twitter dengan menggunakan *k-Nearest Neighbor*(kNN) pada penelitian [1] telah dilakukan penelitian tentang prediksi kepribadian dari *Twitter*, menggunakan metode regressi. Untuk kasus prediksi kepribadian pada peneliti [2] klasifikasi kepribadian berdasarkan teks twitter menggunakan *Naive Bayes*, kNN dan SVM.

Tujuan penelitian ini yaitu untuk mengetahui kepribadian pengguna twitter yang telah di-*survey* dan dilebelkan, maka seberapa besar akurasi yang dihasilkan oleh metode kNN dan TF-IDF yang diaplikasikan pada analisis *Big five Personality*.

## 2. Studi Terkait

Berikut ini studi yang terkait dengan topik tugas akhir ini.

### 2.1 Kepribadia *Big Five*

Kepribadian adalah salah satu ciri yang paling dominan keluar dari diri seseorang, dapat berupa tingkah laku atau sifat dari orang tersebut. Dari kepribadian seseorang kita dapat mengetahui bagaimana pola pikir seseorang untuk mengambil keputusan,bereaksi terhadap sesuatu. Intinya, kepribadian cenderung lebih terbentuk melalui interaksi sosial[3]. Contoh *Big Five* pada perub-ahan perilaku pada Tabel 1.

Tabel 1. *Big five* pada perubahan perilaku

| Skala Trait | Karakteristik skor tinggi | Karakte-ristik skor rendah |
|---|---|---|
| ***Extraversion*** Mengukur kuantitas dan intensitas dari interaksi interpersonal, tingkatan aktifitas, kebutuhan akan dorongan, dan kapasitas dan kesenangan. | Mudah menyesuaikan diri dengan lingkaran sosial, aktif, banyak bicara, orientasi pada hubungan sesama, optimis, *fun loving*, *effectionate*. | Tidak ramah, bersahaja, suka menyendiri, orientasi pada tugas, pendiam. |
| ***Agreeableness*** Mengukur kualitas dari apa yang dilakukan dengan orang lain dan apa yang dilakukan terhadap orang lain. | Lembut hati, dapat dipercaya, suka menolong, pemaaf, penurut. | Sinis, kasar, curiga, tidak kooperatif, pendendam, kejam, manipulative. |
| ***Neuroticism*** Menggambarkan stabilitas emosional dengan cakupan-cakupan perasaan neg-atif yang kuat termasuk kecemasan,kesedi han, irritabilitas dan ketidak percaya diri. | Tenang, santai, merasa aman, puasa terhadap dirinya, tidak emosional, tabah | Cemas, gugup, emosional, merasa tidak aman, merasa tidak mampu, mudah panik. |
| ***Openness to Experience*** Gambaran keluasan, keda-laman,dan kompleksitas mental individu dan peng-alamannya. | Ingin tahu, minat luas, krea-tif, original, im-ajinatif, untraditional. | Konvension al, sederhana, minat sempit, tidak *artistic*, dan tidak analitis. |
| ***Conscientiosness*** Mengukur tingkat keter-aturan seseorang,ketahan an dan motivasi dalam mencapai tujuan berlawanan dengan ketergantungan, dan kecende-rungan untuk menjadi malas dan lemah | Teratur, dapat dipercaya, pekerjaan keras, disiplin, tepat waktu, teliti, rapi, ambisius, dan | Tidak bertujuan, tidak dapat dipercaya, malas, kurang perhatian, lalai, sembrono, tidak disiplin, keinginan lemah, suka bersenang-senang. |

Skala lima faktor kepribadian dalam penelitian ini menggunakan skala *Big Five inventory* (BFI) [4]. BFI terdiri dari 44 butir pernyataan dengan reliabilitas masing-masing ciri sebesar 0.685(*Extrovertion*), 0.677 (*Agreetableness*), 0.461 (*Conscientiosnes*), 0.697 (*Neuroticsm*), 0.704 (*Openness to Experience*).

### 2.2 Penggunaan Fitur

Seperti yang sudah dijelaskan, pada penelitian ini penulis melakukan analisis yaitu pedekatan terhadap perilaku sosial pengguna twitter dan penggunaan bahasa atau kata yang digunakan pengguna twitter pada saat menuliskan *tweet.*

2.2.1 Kepribadian Berdasarkan Perilaku Soisial

Perilaku sosial mendefinisikan kepribadian melalui frekuensi penggunaan media sosial dan tingkat keaktifan antara pengguna. Fitur yang menunjukan tikat perilaku sosial pengguna *Twitter* berdasarkan penelitian yang dilakukan [3] adalah sebagai berikut.

a. *Followers : Follower* adalah penggunaan Twitter lain yang mengikuti pengguna yang di acu.
b. *Following : Following* adalah pengguna yang diacu menjadi *followers* dari pengguna lain.
c. Jumlah *mention: Mention* yang ditandai dengan '@username' menunjukkan tingkat interaksi penggunaan Twitter dengan pengguna lain.
d. Jumlah *hashtag: Hashtag* menunjukkan keterlibatan pengguna dengan isu/topik yang sedang dibahas. *Hashtag* ditandai dengan karakter '#'.
e. Jumlah replay : *Replay* adalah *mention* dari pengguna lain kepada pengguna twitter yang diacu.
f. Jumlah URL : URL adalah tautan berupa informasi website/blog yang dicantumkan pengguna.
g. Jumlah kata : Tweet adalah tulisan yang terdiri dari kumpulan data dengan panjang maksimal 140 karakter dalam *tweet.* Jumlah kata dalam *tweet* adalah total kata yang menyusun *tweet* itu

Selain fitur diatas, terdapat fitur dari Twitter yang dapat dijadikan bahan pertimbangan untuk dilakukan analisis terkait fitur yang menunjukan tingkat keaktifan perilaku sosial pengguna Twitter dari pengguna lainya.

a. Jumlah *retweet* :Retweet adalah tautan berupa gambaran atau video yang di unggah oleh pengguna.
b. Jumlah media URL: Media URL adalah tautan berupa video atau gambar yang digunakan pengguna.
c. Jumlah tanda baca: Tutan berupa simbol dari sebuah kata yang ingin diungkapkan oleh pengguna, tanda baca yang dihitung adalah tanda Tanya '?' dan tanda Seru '!'.
d. Jumlah emoji : Emoji adalaj karakter unik yang dapat digunakan oleh pengguna saat menuliskan tweetnya untuk menggambarkan emosi pengguna melalui karakter-karakter unik. Emoji yang diambil dari *link* dan di simpan kedalam kamus pada database. Total emoji yang didapatkan berjumlah 2.552 karakter.
e. Rata-rata kata : Rata rata kata adalah jumlah dari kata yang dituliskan pengguna di *teet* dibagi dengan jumlah *tweet* yang berhasil di crawling dan sudah dikurang dengan jumlah *retweet* karena *retweet* uka kata yang ditulis oleh pengguna.
f. Jumlah huruf besar : Huruf besar adalah huruf kapita atau simbol-simbol yang menyusun sebuah *tweet.*
*g.* Jumlah karakter : Karakter adalah susunan huruf atau simbol-simbol yang menyusun sebuah *tweet.*
h. Rata-rata karakter : Rata rata karakter adalah jumlah dari karakter yang dituliskan pengguna di *tweet* d bagi dengan jumlah *tweet* yang erhasil di *crawling* dan sudah dikurang dengan jumlah *retweet* karena *retweet* bukan karakter yang dituliskan langsung oleh pengguna.

2.2.2 Kepribadian Berdasarkan Linguistik

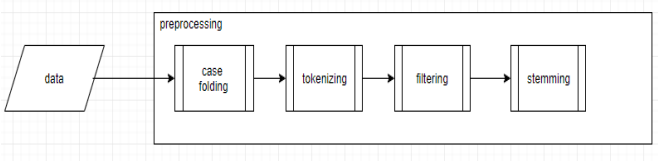
Pada pendekatan Linguistik, fitur atau atribut melakukan pencarian sendiri terhadap penggunaan kata di *tweet* yang telah dikumpulkan untuk menemukan hubungan antara kata dengan kepribadian pengguna Twitter. Hasil yang didapatkan dari pendekatan linguistik ini adalah pengetahuan baru mengenai kaitan kata/bahasa dengan kepribadian pengguna Twitter. Fitur linguistik bekerja dengan cara menguraikan *tweet* ke dalam satuan kata dengan pendekatan unigram. Setelah diuraikan, satuan kata tersebut diberi bobot dengan perhitungan TF-IDF.

2.3 *Pre-processing*

*Pre-processing* adalah suatu proses pengubahan bentuk data yang belum terstruktur menjadi data yang terstruktur sesuai dengan kebutuhan, untuk proses *mining* yang lebih[5]. Metode yang sangat penting dalam melakukan teks *mining*, ini adalah langkah pertama dalam *text mining, pre-processing* dilakukan untuk membuat data mentah menjadi data yang berkualitas, berikut ini tahapan *pre-processing ;*

a. *Case folding* : mengubah semua huruf menjadi huruf kecil.
b. *Tokenizing* : tahapan pemotongan string berdasarkan tiap kata.
c. *Filtering* : tahap mengambil kata-kata penting dari hasil *tokenizing.*
d. *Stemming :* merupakan suatu proses yang terdapat dalam sistem IR (*Information Retrive*) yang mentransformasikan kata kata yang terdapat dalam suatu dokumen ke kata kata.

Tahapan *pre-processing* yang berupa kalimat dari setiap akun di pecah menjadi per kata dapat dilihat pada gambar 1.




Gambar 1. Proses *Pre-Processing*

2.4 *Term-Frequency inverse Document Frequency (TF_IDF)*

*Term-frequency inverse Document frequency* merupakan salah satu metode pembobotan yang menggabungkan antara *Term-Frequency(TF)* dan *Inverse Document Frequency (IDF)* atau pembobotan pada setiap kata dalam setiap dokumen teks [6]. *Term-Frequency* Adalah frekuensi dari kemunculan sebuah kata dalam dokumen yang bersangkutan sedangkan *Inverse Document Frequency* adalah banyaknya dokumen yang mengandung kata tertentu.

$$IDF(w) = \log\left(\frac{N}{DF(w)}\right) \quad (1)$$

TF-IDF(w,d)=TF(w,d)xIDF(w) (2)

Keterangan;
TF-IDF(w,d) : bobot kata dalam sebuah dokumen
W : suatu kata
D : dokumen
TF(w,d) : frekuensi kemunculan w dalam d
IDF : inverse DF dari w
N : banyaknya kelas
DF(w) : banyaknya kata dalam kelas

2.6 *k-Nearest Neighbor (kNN)*

Klasifikasi merupakan sebuah metode untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan dapat memperkirakan kelas dari suatu objek yang lebelnya tidak diketahui. Model itu sendiri bisa berupa aturan "jika-maka", berupa *decision tree*, formula matematis atau *neural network.* Metode-metode klasifikasi antara lain *RainForest, Naive Bayesiann, Neural Network, Genetic Algorthm, Fuzzy, Case-based Ressoning* dan *k-Nearest Neighbor* [7].

*k-Nearest Neighbor (*kNN*)* adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (*train dataset*), diambil dari k tetangga terdekatnya (*Nearest Neighbors*), dengan k merupakan banyaknya tetangga[6]. Salah satu metode yang menerapkan algoritma *supervised*. Perbedaan antara *supervised learning* dengan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang suda ada, dan sedangkan *unsupervised learning*, data belum memiliki pola apapun[8]. Berdasarkan kategori pada algoritma kNN, dimana kelas yang paling banyak muncul nantinya akan menjadi kelas dari hasil klasifikasi. Berikut merupakan langkah-langkah algoritma kNN :

a. menentukan parameter k (jumlah tetangga terdekat),
b. menghitung jarak objek terhadap data latih yang diberikan,
c. mengurutkan hasil no 2 secara *ascending* (dari nilai tinggi ke rendah),
d. mengumpulkan kelas (klasifikasi *Nearest Neighbor* berdasarkan nilai k) dan,
e. dengan menggunakan kategori *Nearest Neighbor* yang paling mayoritas maka dapat diprediksikan sebagai kelas objek.

Untuk mengidentifikasi jarak antara dua titik yaitu pada data latih *(x)* dan titik pada data uji *(y)* digunakan rumus *euclidean distance.*

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

Keterangan
D :jarak antara titik
*x* *:data training (*data latih*)*
*y* *:data testing* (data uji)

2.7 *Conffusion Matrix*

*Confusion matrix* merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi[9]. Metode ini memiliki 4 keluaran yaitu pada tabel 2.

Tabel 2. *Confussion matrix*

| *Actual* / *Prediction* | *Positive* | *Negative* |
|---|---|---|
| *Positive* | *True Positive (tp)* | *False Positive (fp)* |
| *Negative* | *False Negative (fn)* | *True Negative (tn)* |

*a. Recall*

*Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali seuah informasi, didefinisikan dengan

$$Recall = \frac{tp}{tp+fn} \quad (4)$$

*b. Precission*

*Recision* adalah presentase dari nilai yang di prediksi *true* dan terbukti *true*, didefinisikan dengan

$$Precission = \frac{fp}{fn+fp} \quad (5)$$

*c. Accuracy*

*Accuracy* adalah sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual, didefinisikan dengan :

$$Accuracy = \frac{tp+tn}{tp+fn+tn+fp} \quad (6)$$

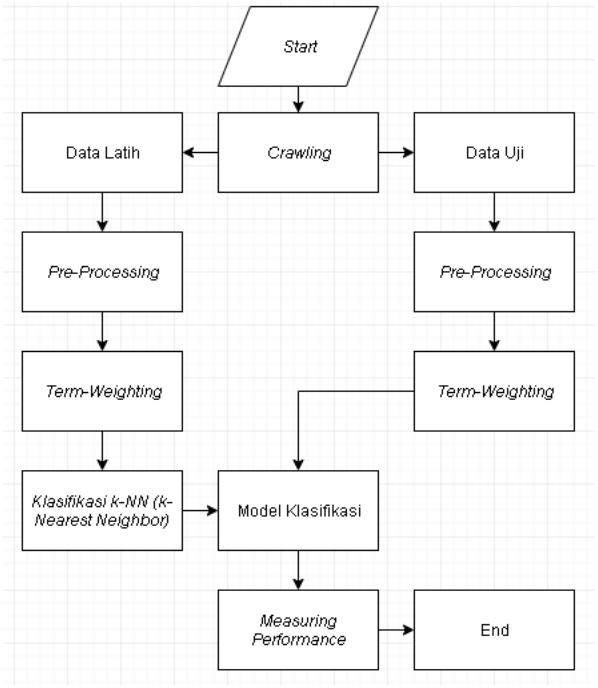
Keterangan:

***tp*** : hasil prediksi positif dan data sebenarnya positif
***tn***: hasil prediksi negatif dan data sebenarnya negatif
***fp*** : hasi prediksi negatif dan data sebenarnya negatif
***fn*** : hasil prediksi negatif dan data sebenarnya positif

Sesuai dengan tujuan pada penelitian kali ini yaitu menghitung akurasi yang dihasilkan oleh metode kNN maka digunakan perhitungan *confussion matrix* untuk menghitung *accuracy.*

## 3. Sistem yang Dibangun

Sistem umum yang akan dibuat dalam penelitian ini bisa dilihat pada gambar 2.




Gambar 2. Sistem Klasifikasi Menggunakan kNN

### 3.1. *Crawling* Data

*Crawling data* adalah proses mengambil data dari sebuah *website Twitter* menggunakan *Twitter* API. Atribut yang diambil merupakan atribut kebutuhan dari fitur yang akan dianalisis. Setiap *crawling,* data *tweet* yang diambil di setiap akunnya bisa sampai 3200 *tweet* terbarunya. Dari kuisioner yang penulis bagikan terdapat 228 responden yang mengisi tetapi akun yang dapat di-*crawling* hanya sejumlah 137 akun, dari 137 akun yang terkumpul mesin *crawling* dapat mengumpulkan data sebanyak 331.439 *tweet.* Mengandung 85 kelas O*penness to Experience,*16 *Conscientiousness,* 14 *Extraversion, 10 Agreableness,* dan *12 Neuroticsm.*

### 3.2 Pembagian data

Pada tahap ini data yang sudah dikumpulkan akan dipisahkan menjadi dua data yaitu data lati dan data uji . jumlah rasio pada penelitian ini penulis membuat skenario pembagian data menjadi (70:30)

### 3.3 Pelabelan

Pada proses pelabelan dilakukan pemberian label kelas atau pengelompokan kelas sesuai jenis kepribadian berdasarkan *big five.* Caranya adalah memberi kuisoner test kepribadian kepada sampel yang suka rela dan akan melibatkan pakar untuk memeriksa dan memberikan label pada sampel tersebut.

### 3.4 *Pre-processing*

Pada *Pre-processing* ini bertujuan untuk mengubah bentuk data acuan yang tidak tersetruktur menjadi bentuk terstruktur sesuai dengan kebutuhannya. *Pre-processing* data yang dilakukan terhadap data teks *tweet* adalah *Case Foldeing, Tokenizing, Filtering, Stemming.*

### 3.5 Pembobotan

Pada proses ini masing-masing data yang sudah selesai di *pre-processing* diberikan nilai atau bobot dengan menggunakan metode TF-IDF.

Pembobotan ini berguna untuk mengukur seberapa pengaruh kata dari suatu dokumen nantinya

3.6 *Term-Weighting*

Pada proses ini masing masing data *tweet* yang sudah selesai di *pre-processing* akan diberikan nilai atau bobot dengan menggunakan pembobotan TF-IDF dengan pendekatan fitur *Bag of Word* yang didukung dengan ekstraksi fitur unigram. Pembobotan ini berguna untuk mengukur seberapa berpengaruh kata dari suatu dokumen *tweet* nantinya. *Term* yang di pilih untuk dijadikan atribut ditentukan kedalam beberapa skenario pada 50 kata tersering muncul menjadi 10.

3.7 Klasifikasi dengan *k-Nearest Neighbor*(kNN)

Pada proses ini data yang telah di-*pre-processing* akan masuk ke tahap klasifikasi. Dimana data latih akan diinput dan dihitung berdasarkan proses *k-Nearest Neighbor.* Output dari proses ini adalah model prediksi yang nanti akan diujikan performansinya.

3.8 Model Prediksi

Model prediksi adalah sistem pembelajaran yang sudah dibuat dari klasifikasi kNN. Data uji yang sudah dibua skenarionya akan diuji terhadap model yang telah dibuat. Output adalah nilai performansi dari model yang dibuat.

3.9 Pengaruh Parameter Nilai k

Proses pengujian kemudian dilanjutkan dengan melihat pengaruh nilai k terhadap skenario yang memiliki akurasi yang cukup tinggi

**4. Hasil dan Analisis**

Skenario yang telah dibuat akan diuji menggunakan data latih yang telah di dapatkan dari tahap pemecahan data.

4.1 Data Set dan Pelabelan

Sebelum melakukan pemecahan data, data terlebih dahulu dilabeli secara manual, contoh pelabelan secara manual dapat dilihat pada Data akun *Twitter* yang digunakan pada penelitian ini sebanyak 137 orang yang berasal dari *crawling twitter*. Data tersebut kemudian dilabelkan secara manual ke dalam 5 kelas yaitu *Openness to Experience, Conscientiosness, Extraversion, Aggre-eableness,* dan *Neuroticsm.*

4.2 Hasil Pengujian

Berikut ini merupakan hasil pengujian dan klasifikasi dari sistem yang telah dibuat

4.2.1 Pemecahan Data

Sebelum masuk fitur dasar dan fitur pengaruh TF-IDF data di uji terlebih dahulu untuk mengetahui komposisi data yang paling baik untuk fitur selanjutnya, setiap sel dilakukan pengujian secara acak, diketahui bahwa data akurasi dengan pemecahan data 70%:30% memiliki akurasi sebesar 60,97%

4.2.2 Perhitungan dan Pembobotan

Pada tahap ini dilakukan *pre-processing* dan pembobotan dengan perhitungan TF-IDF. Dimana kata pada setiap *tweet* di akun 1,2,3...,n akan dihitung berapa banyak kemunculan TF(*Term-Frequencey*). Selanjutnya nilai dari TF 1,2,3,...,n akan dikalikan dengan hasil perhitungan IDF. Nilai dari hasil pengalian pada akun TF-IDF 1,2,3,...,n menghasilkan bobot untuk setiap kata yang akan di proses kembali untuk ke tahap selanjutnya yaitu penghitungan metode *k-Neares Neighbor.*berikut contoh tabel kata TF-IDF pada Tabel 3.

Tabel 3. Contoh tabel kata hasil TF-IDF

| No | Akun | *Document(0,C,E,A,N)* | | | |
|---|---|---|---|---|---|
| | | manta p | kuas a | gun a | ... |
| 1 | TF aderizkyputr ii | 13 | 3 | 4 | ... |
| 2 | TF 12290F | 0 | 0 | 3 | ... |
| 3 | TF alyalarasatiii i | 0 | 0 | 0 | ... |
| ... | .. | .. | .. | .. | .. |
| 137 | TF yasinfjr | 1 | 1 | 0 | .. |
| 1 | TF_IDF aderizkyputr ii | 5,13 | 1,75 | 0,76 | ... |
| 2 | TF _IDF 12290F | 0,00 | 0,00 | 0,57 | ... |
| 3 | TF_IDF alyalarasatiii i | 0,00 | 0,00 | 0,00 | ... |
| ... | .. | .. | .. | .. | .. |
| 137 | TF_IDF yasinfjr | 0,27 | 0,58 | 0,00 | .. |

4.2.3 kNN(*k-Nearest Neighbor*).

Selanjutnya pengerjaan metode kNN, data dipisahkan menjadi dua, yaitu data latih dan data *uji* kemudian dilatih, untuk mendapatkan *clasifier* untuk menebak kelas prediksi yang langsung dimasukan ke dalam sistem yang telah dibuat.

4.2.4 Pengaruh Perilaku Sosial dan TF-IDF

Proses pengujian dengan pengaruh data sosial dan linguistik sangatlah berpengaruh terhadap hasil akurasi. Dengan membandinganken hasil label dari sistem dengan hasil label dari pelabelan manual sehingga didapatkan akurasi terbaik dari 19 Skenario yang dijalankan. Berikut data hasil pengaruh perilaku sosial dan linguistik pada tabel 4

Tabel 4. Pengaruh PS dan Linguistik

| Skenario | Fitur yang Digunakan | Akurasi (%) |
|---|---|---|
| Perilaku Sosial (PS) | *Followers* | 48,7805 |
| | *Followers*+*Following* | 43,9024 |
| | *Followers*+*Following* +MediaURL+URL | 51,2195 |
| | *Followers*+*Following* +T.Mention | 53,6585 |
| | Media URL+*Mention*+RT+ Huruf Besar+T.Kata | 29,2683 |
| | *Mention*+RT+Huruf Besar | 36,5854 |
| | *Followers*+*Following* +Tanda Baca+Emoji | 51,2195 |
| | **Rata2 Kata+T.karakter** | **58,5366** |
| | T.Kata+Emoji | 46,3415 |
| | T.RT+Huruf Besar+T.Karakter | 34,1463 |
| Pengaruh TF-IDF | *Followers*+*Following* +TF_IDF | 48,7805 |
| | *Followers*+*Following* +MediaURL+URL+ TF_IDF | 48,7806 |
| | *Followers*+*Following* +T.Mention+TF_IDF | 48,7807 |
| | Media URL+*Mention*+RT+ Huruf Besar+T.Kata+TF_IDF | 39,0244 |
| | *Mention*+RT+Huruf Besar+TF_IDF | 39,0245 |
| | *Followers*+*Following* +Tanda Baca+Emoji+TF_IDF | 46,3415 |
| | Rata2 Kata+T.karakter+TF_ IDF | 46,3415 |
| | T.Kata+Emoji+TF_IDF | 46,3416 |
| | T.RT+Huruf Besar+T.Karakter +TF_IDF | 39,0244 |

4.2.5 Pengukuran Performansi

Proses pengujian dengan melihat pengaruh nilai k terhadap memiliki akurasi tertinggi 60,97%. Berikut pengaruh nilai k terhadap performansi pada tabel 5.

Tabel 5. Pengaruh Nilai K

| Nilai K | Akurasi % |
|---|---|
| 1 | 39,0244 |
| 2 | 46,3415 |
| 3 | 48,7805 |
| 4 | 48,7805 |
| 5 | 53,6585 |
| 6 | 58,3333 |
| 7 | 58,5366 |
| 8 | 60,0000 |
| **9** | **60,9756** |
| 10 | 60,0000 |
| 11 | 58,5366 |
| 12 | 58,9744 |
| 13 | 57,5000 |
| 14 | 58,5366 |

Berikut merupakan tabel 6. *confussion matrix* berdasarkan skenario terbaik dengan nilai k=11 dari hasil pengujian klasifikasi menggunakan kNN.

Tabel 6. Hasil *confussion Matrix*

| | | PREDICT | | | | |
|---|---|---|---|---|---|---|
| | | O | C | E | A | N |
| ACTUAL | O | 24 | 0 | 0 | 0 | 0 |
| | C | 5 | 0 | 0 | 0 | 0 |
| | E | 5 | 0 | 0 | 0 | 0 |
| | A | 1 | 1 | 0 | 1 | 0 |
| | N | 4 | 0 | 0 | 0 | 0 |

Dari hasil pengujian *confussion matrix* didapatkan *recall* sebesar 61.53% pada label ‘O’,100% pada label ‘A’ dan pada *precision* sebesar 100% pada label ‘O’ , 33,33% pada label ‘A’

4.4 Analisis Hasil Pengujian

Dari hasil percobaan menggunakan kNN (*K-Nearest Neighbor*)dan TF-IDF dengan rasio data set 70%:30%, didapatkan akurasi sebesar 60,97%. Penelitian ini menggunakan bobot fitur *Mention*, RT, total *hastag*, Media URL, total *Followers, total Following*, total URL, dan TF-

IDF. Penelitian ini juga masih menggunakan basis kata unigram, dimana dalam tahap pengujian belum maksimal untuk melakukan uji similitaritas data.

**5. Kesimpulan**

Setelah melakukan penelitian prediksi kepribadian *Big five* menggunakan TF-IDF dan dengan metode k-NN ini dapat ditarik kesimpulan ,yaitu akurasi tertinggi dari komposisi perilaku soial dan linguistik dan dengan pengukuran performansi dengan melihat nilai k=9 sebesar 60,97%, sedangkan komposisi perilaku sosial dan linguistik dengan performansi nilai k=1 memiliki nilai terendah dengan nilai 39,02%. Meskipun sudah melakukan *item* kuisoner berdasarkan jumlah yang sama di setiap dimensi dengan nilai validasi dan reliabilitas yang baik, hasil skoring tetap mendapatkan label yang lebih banyak di satu dimensi kelas yaitu *Opennes to Experience* . Sehingga, kecil nilai akurasi dikarenakan jumlah data latih yang lebih dominan banyak di kelas *Openness to Experience* yaitu sebanyak 61 data latih dan 24 di data uji, sehingga model prediksi yang dibangun cenderung memprediksikan setiap keputusan adalah *Openness to Experience* dan membuat nilai *precision* dan *recall* pada *Openness to Experience* meningkat tetapi tidak pada kelas lainya.

Fitur perpaduan pendekatan perilaku pengguna dengan pendekatan linguistik menggunakan pembobotan TF-IDF dengan system unigram, akurasi yang didapatkan tidak sesuai dengan yang diharapkan sebesar 70%. Ketimpangan suatu data pada kelas tertentu agar keputusan pada saat membuat model prediksi tidak cenderung kepada data kelas yang dominan.

## Daftar Pustaka


[1] S. Adali and J. Golbeck, "Predicting Personality with Social Behavior," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2012.

[2] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," *Proc. 2015 Int. Conf. Data Softw. Eng. ICODSE 2015*, pp. 170–174, 2016.

[3] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," in *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 2011.

[4] S. S. John, O. P., "Big Five Inventory ( Bfi )," in *Handbook of Personality Second Edition: Theory and Research*, vol. 2, 1999, pp. 102–138.

[5] A. R. Naradhipa and A. Purwarianti, "Sentiment classification for Indonesian message in social media," in *Proceedings - International Conference on Cloud Computing and Social Networking 2012: Cloud Computing and Social Networking for Smart and Productive Society, ICCCSN 2012*, 2012.

[6] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," in *Procedia Engineering*, 2014, vol. 69, pp. 1356–1364.

[7] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter Profiles,Our Selves: Predicting Personality with Twitter."

[8] N. Krisandi, B. Prihandono, and Helmi, "Algoritma K - Nearest Neighbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada PT. MINAMAS Kecamatan Parindu," *Bul. Ilm. Math.Stat.dan Ter.*, vol. 02, no. 1, pp. 33–38, 2013.

[9] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.