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)

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana dari

Program Studi Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

1107120013

FAIDH ILZAM NUR HAQ

Program Studi Sarjana Ilmu Komputasi

Fakultas Informatika

Universitas Telkom

Bandung

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)

Faidh Ilzam Nur Haq[1], Erwin Budi S, S.Si., M.T. [2]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[4]Divisi Digital Service PT Telekomunikasi Indonesia
[1]filzam@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id,

Abstrak

Analisis kepribadian seseorang sangat membantu sebagai penilaian dalam berbagai hal seperti perekrutan, karir, dan 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 informasi data pengguna twitter tersebut, hal ini untuk menambah metode dari analisis kepribadian. Teori klasifikasi kepribadian menggunakan *Big Five Personality* yang terdiri dari *openness, conscientiousness, extraversion, agreeableness,* dan *neuroticism.* Metode yang digunakan yaitu *Naïve Bayes Classifier* dengan pembobotan menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Term Frequency-Relevance Frequency* (TF-RF).

Kata Kunci : Twitter, *Big Five Personality, Naïve Bayes Classifier, TF-IDF* dan *TF-RF*.

Abstract

Analysis of a person's personality is helpful as an assessment in such matters as hiring, career, and health. The usual method used in personality analysis is by interview, observation, and questionnaire survey. This research tries to give solution with enough use social media that is twitter, by analyzing twitter user data information, this is to add method from personality analysis. Personality classification theory uses Big Five personality consisting of openness, conscientiousness, extraversion, agreeableness, and neuroticism. The method used is Naïve Bayes Classifier with weighting using Term Frequency-Inverse Document Frequency (TF-IDF) *and Term Frequency-Relevance Frequency* (TF-RF).

Keywords : *Twitter, Big Five Personality, Naïve Bayes Classifier, TF-IDF and TF-RF.*

1. Pendahuluan

Kepribadian merupakan keseluruhan bentuk perilaku seseorang individu yang dipunyai sebagai latar belakang pemiliknya terhadap lingkungannya. Kepribadian dapat mempengaruhi pilihan hidup seseorang dalam berbagai hal dan mempengaruhi interaksi dengan orang lain dan lingkungan. Ada banyak metode klasifikasi kepribadian salah satunya yang banyak digunakan yaitu metode *Big five personality* yang terdiri dari *openness, conscientiousness, extraversion, agreeableness,* dan *neuroticism*[1]. Kepribadian dapat digunakan sebagai penilaian dalam berbagai hal seperti perekrutan, karir, dan kesehatan dengan melakukan tes kepribadian untuk mengetahui kepribadiannya[2]. Banyak metode untuk melakukan tes kepribadian dengan cara wawancara, observasi, atau survei kuesioner, namun metode ini masih kurang praktis dan mahal. Dalam sebuah studi baru-baru ini menunjukan bahwa tes kepribadian dapat diperoleh dari media sosial dengan menganalisa teks yang mereka tulis[3][4][5], metode tes kepribadian ini lebih praktis dan murah.

Media sosial merupakan tempat seseorang untuk berinteraksi dengan individu dalam dunia maya, akun media sosial ini bersifat pribadi sehingga mencerminkan kepribadian seseorang tersebut. Salah satu media sosial yang banyak digunakan yaitu Twitter dengan jumlah pengguna 328 juta pada tahun 2017 di seluruh dunia[6]. Penelitian tugas akhir ini menggunakan Twitter dikarenakan banyak digunakan dan pengguna twitter cenderung mencurahkan pendapatnya lewat kata-kata yang dibatasi 280 karakter sehingga lebih langsung menyampaikan maksud pesan. Kata-kata bisa menggambarkan kepribadian seseorang dengan menganalisa untuk mendapatkan informasi dari pengguna Twiiter. Metode yang digunakan untuk menganalisa kata menggunakan *Linguistic Inquiry and Word Count* (LIWC) yaitu program analisis teks di bidang ilmu psikologi[12].

Dalam penelitian tugas akhir ini mencoba melakuan prediksi kepribadian *big five* melalui Twitter dengan menggunakan algoritma *Naïve Bayes Classifier* dikarenakan implementasi yang sederhana untuk dibangun dengan waktu komputasi yang relatif cepat dan tingkat akurasi yang tinggi dibandingkan dengan algoritma yang lain[4]. Hal lain yang membedakan dengan penelitian [4] yaitu menggunakan kuisuner *Big Five Inventory* (BFI)[1[dan menggunakan pembobotan nilai *Term Frequency-Inverse Document Frequency* (TF-IDF) dan

Term Frequency-Relevance Frequency (TF-RF) untuk membandingkan nilai akurasi kedua pembobotan tersebut.

2. Studi Terkait

2.1. Media Sosial

Dalam perkembangannya media sosial selalu mengikuti perkembangan teknologi yang cepat, banyak sekali saat ini media sosial yang digunakan seperti *Facebook, Twitter, Instagram, Path, Line, WhatsApp* dan masih banyak lagi. Salah satu media sosial yang besar sampai saat ini yaitu Twitter, berdasarkan data PT Bakrie Telecom jumlah pengguna Twitter sampai 19,5 juta pengguna di Indonesia dari total 500 juta pengguna global[6]. Pada penelitian menggunakan media sosial twitter yang merupakan sebuah layanan jejaring sosial dan mikroblogong daring yang masih banyak digunakan hingga saat ini. Twitter didirikan oleh Jack Dorsey pada tahun 2006 yang digunakan pertama kali untuk layanan internal Odeo dan sekarang dioperasikan oleh Twitter, Inc. Pengguna Twitter dapat mengirim dan membaca teks yang diunggah dengan maksimal jumlah karakter 280 setelah dikembangkan dari 140 karakter yang disebut dengan *tweet*.

Kebiasaan pengguna Twitter sering memplubikasikan kegiatan sehari-hari maupun mencurahkan pendapatnya yang dapat berupa kata-kata maupun foto-foto, hal ini dapat digunakan untuk memprediksi kepribadian pengguna Twitter dari *tweet* berisi data pribadi yang diunggah[8].Twitter telah menyediakan *Application Programming Interface* (API) yang berguna untuk memudahkan para *developer* atau pengembang mengambil data dari Twitter untuk diolah. Twitter API merupakan sekumpulan perintah, fungsi, komponen dan protokol yang disediakan Twitter yang dapat digunakan untuk mengembangkan oleh pihak lain. Dengan fitur API Twitter dapat digunakan untuk prediksi karakter pengguna Twitter dengan mengambil data Twitter yang berisi informasi mengenai *tweet* dan data pengguna[9].

2.2. Big Five PersonalityTraits Model

Tes psikologi merupakan tahapan tes yang sering digunakan untuk sebagai pertimbangan dalam merekrut seseorang untuk diterima maupun ditempatkan pada instansi manapun. Ketidaksesuaian kepribadian seseorang terhadap pekerjaan yang ditugaskan akan mengakibatkan kerugian bagi instansi maupun pekerja itu sendiri.

Banyak penelitian dan teori tentang sifat kepribadian dan paling sering digunakan dalam dunia kerja adalah teori sifat kepribadian "Model Lima Besar" atau "*Big Five Personality Traits Model"* yang dikemukan oleh Lewis Goldberg[1]. Teori model lima besar ini terdiri dari 5 dimensi yaitu *Openness, Conscientiousness, Extraversion, Agreeablenesss,* dan *Neuroticism*[10]. Berikut adalah penjelasan mengenai Sifat Kepribadian Model Lima Besar :

1. *Openness to Experience* (Terbuka terhadap Hal-hal baru)

Dimensi Kepribadian *Opennes to Experience* ini mengelompokan individu berdasarkan ketertarikannya terhadap hal-hal baru dan keinginan untuk mengetahui serta mempelajari sesuatu yang baru. Individu *openness* mempunyai karakteristik positif yaitu cenderung lebih kreatif, imajinatif, intelektual, penasaran dan berpikiran luas. Individu dengan tingkat *openness* rendah cenderung konvensional dan nyaman terhadap hal-hal yang telah ada serta akan menimbulkan kegelisahan jika diberikan tugas-tugas baru.

2. *Conscientiousness* (Sifat Berhati-hati)

Dimensi Kepribadian *conscientiousness* ini cenderung lebih berhati-hati dalam melakukan suatu tindakan ataupun penuh pertimbangan dalam mengambil sebuah keputusan, mereka juga memiliki disiplin diri yang tinggi dan dapat dipercaya. Individu *conscientiousness* mempunyai karakteristik positif yaitu dapat diandalkan, bertanggung jawab, tekun dan berorientasi pada pencapain. Sifat kebalikan dari *conscientiousness* adalah individu yang cendurung kurang bertanggung jawab, terburu-buru, tidak teratur dan kurang dapat diandalkan dalam melakukan suatu pekerjaan.

***3.* *Extraversion* (Ekstraversi)**

Dimensi Kepribadian *extraversion* ini berkaitan dengan tingkat kenyamanan seseorang dalam berinteraksi dengan orang lain. Karakteristik Positif Individu *extraversion* adalah senang bergaul, mudah bersosialisasi, hidup berkelompok dan tegas. Sebaliknya individu yang *introversion* (kebalikan dari *extraversion*) adalah mereka yang pemalu, suka menyendiri, penakut dan pendiam.

4. *Agreeableness* (Mudah Akur atau Mudah Bersepakat)

Dimensi *agreableness* ini cenderung lebih patuh dengan individu lainnya dan memiliki kepribadian yang ingin menghindari konflik. Karakteristik positif *agreableness* adalah kooperatif (dapat bekerjasama), penuh kepercayaan, bersifat baik, hangat dan berhati lembut serta suka membantu. Karakteristik kebalikan dari sifat *Agreeableness* adalah mereka yang tidak mudah bersepakat dengan individu lain karena suka menentang, bersifat dingin dan tidak ramah.

5. *Neuroticism*

Neuroticism adalah dimensi kepribadian yang menilai kemampuan seseorang dalam menahan tekanan atau stress. Karakteristik Positif dari Neuroticism disebut dengan *Emotional Stability* (Stabilitas Emosional), Individu dengan Emosional yang stabil cenderang tenang saat menghadapi masalah, percaya diri, memiliki pendirian yang teguh. Sedangkan karakteristik kepribadian *neuroticism* (karakteristik negatif) adalah mudah gugup, depresi, tidak percaya diri dan mudah berubah pikiran.

Pengukuran kepribadian *Big Five Personality Traits Model* ini dapat dilakukan dengan menggunakan berbagai metode, salah satunya metode BFI[1] berupa kuesioner berisi 44 pertanyaan yang dapat menggambarkan kepribadian *Big Five Personality Traits Model*.

2.3. Pengunaan fitur

Pada penelitian ini menggunakan pendekatan berdasarkan perilaku sosial pengguna twitter melaui fitur yang ada pada twitter seperti *follower, following, mention, hashtag, reply, url, tweet, retweet,media url, emoji* selain itu ditambah juga dengan fitur jumlah tanda baca, jumlah huruf besar, jumlah karakter, rata-rata kata, dan rata-rata karakter[5].

Selain pendekatan perilaku sosial penelitian ini juga mengunakan pendekatan linguistik berdasarkan kata-kata yang digunakan dalam menuliskan *tweet* dan akan diberi bobot dengan perhitungan TF-IDF dan TF-RF.

2.4. Linguistic Inquiry and Word Count (LIWC)

Linguistic Inquiry and Word Count (LIWC) adalah sebuah program analisis bahasa yang paling umum digunakan dalam studi menyelidiki hubungan antara penggunaan kata dan variabel psikologis. Kamus LIWC mendefinisikan lebih dari 70 kategori berbeda (misalnya emosi negatif, seksualitas, pekerjaan, dll.), yang sebagian besar berisi beberapa puluh atau ratusan kata. Skor untuk masing-masing kategori dihitung dengan membagi jumlah kemunculan semua kategori tersebut dengan jumlah total kata[12].

2.5. Term Weighting

Term weighting adalah sebuah metode pembobotan kata (*term*) untuk memberikan sebuah bobot atau nilai untuk kata (*term*) yang terkandung dalam sebuah dokumen. Bobot nilai ini menjadi ukuran besarnya jumlah dan tingkat kontribusi sebuah kata (*term*) untuk penentuan suatu kelas atau kategori dalam suatu dokumen [15]. Terdapat beberapa metode pembobotan kata (*term weighting*) diantaranya adalah TF, TF-IDF, WIDF, TF-CHI, dan TF-RF. Dalam penelitian tugas akhir ini mencoba menguji 2 metode pembobotan kata yaitu TF-IDF dan TF-RF untuk membandingkan tingkat perfomansi.

2.5.1 *Term Frequency-Inverse Document Frequency* (TF-IDF)

TF-IDF merupakan gabungan dari metode *Term Frequency* (TF) dengan *Inverse Document Frequency* (IDF) yang dihasilkan dari perkalian metode TF dengan metode IDF. Metode TF-IDF ini memberikan nilai bobot yang tinggi kepada *term* yang sering muncul pada suatu dokumen, tetapi jarang muncul dalam kumpulan dokumen[15]. Persamaan metode TF-IDF sebagai berikut:

$$TF * IDF(d,t) = TF(d,t) * \log\frac{N}{df(t)} \quad (1)$$

Dimana:

- $TF * IDF(d,t)$: Pembobotan TF-IDF.
- $TF(d,t)$: Frekuensi munculnya *term t* pada dokumen d.
- N : Jumlah dari semua kumpulan dokumen.
- $df(t)$: Jumlah dari dokumen yang mengandung *term t*.

2.5.2 *Term Frequency-Relevance Frequency* (TF-RF)

TF-RF adalah metode gabungan antara TF dan RF dengan tujuan untuk mendapatkan performansi yang lebih baik dari metode pembobotan kata lainnya. Metode ini mempertimbangkan relevansi dokumen dilihat dari frekuensi kemunculan *term* di kategori yang berkaitan [15]. Persamaan metode TF-RF sebagai berikut:

$$TF * RF(t,c) = TF(d,t) * \log_2\left(2 + \frac{a}{\max(1,c)}\right) \quad (2)$$

Dimana:

- $TF * RF(d,t)$: Pembobotan TF-RF.
- $TF(d,t)$: Frekuensi munculnya *term t* pada dokumen d.
- c : Kelas kategori.
- t : Term.

2.6. Klasifikasi

Klasifikasi merupakan dua bentuk analisis data yang dapat digunakan untuk mengekstrak model yang menggambarkan kelas data atau untuk memprediksi tren data masa depan dan dapat membantu memberikan pemahaman yang lebih baik tentang data secara luas. Proses klasifikasi dibagi menjadi dua tahap yaitu *learning* dan *test*, pada tahap *learning* data yang diketahui kelas datanya digunakan untuk membanggun model dan tahap *test* dilakukan untuk menguji model yang sudah terbangun untuk mengetahui tingkat akurasi. Banyak metode klasifikasi dan prediksi telah diajukan oleh peneliti dalam pembelajaran mesin, pengenalan pola, dan statistik.

2.7. Naïve Bayes Classifier

Naïve bayes merupakan salah satu metode klasifikasi yang berdasarkan teorema bayes dengan menggunakan teknik probalitas dan statistik. Algoritma *Naïve Bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya dengan ciri utama adalah asumsi yang sangat kuat akan independensi dari masing-masing kondisi[14]. Kelebihan *Naïve Bayes* adalah tidak membutuhkan jumlah data pelatihan (*data training)* yang banyak untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. *Naïve Bayes* menghitung peluang masuknya sampel karakteristik tertentu dalam kelas c (*posterior*) yaitu peluang munculnya kelas c dikali dengan peluang kemunculan karakteristik sampel pada kelas c (*likelihood)*, dibagi dengan peluang kemunculan karakteristik sampel secara global*(evidance)*. Secara matematis rumus persamaan *Naïve Bayes* sebagai berikut:

$$P(C|X) = \frac{(P(x|c) \text{ x } P(c)}{P(x)} \quad (3)$$

Dimana:

$\boldsymbol{x}$: Data dengan kelas yang belum diketahui
$\boldsymbol{c}$: Data testing yang merupakan kelas-kelas hasil klasifikasi
$\boldsymbol{p(c|x)}$: Peluang hipotesis *c* berdasarkan kondisi *x (Posterior Probability)*
$\boldsymbol{P(x|c)}$ **:** Peluang berdasarkan kondisi *x* pada hipotesis *c (Likelihood)*
$\boldsymbol{p(c)}$ **:** Peluang hipotesis *c (Class Prior Probability)*
$\boldsymbol{P(x)}$ **:** Peluang *x (Predictor Prior Probality)*

Untuk klasifikasi dengan data kontinyu digunakan rumus *Densitas Gauss*[7] :

$$P(X_i = x_i|Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma ij}} e^{-\frac{(x_i - \mathrm{u}_{ij})^2}{2\sigma^2 ij}} \quad (4)$$

Dimana:

P : Peluang
$\boldsymbol{X_i}$: Atribut ke i
$\boldsymbol{x_i}$: Nilai atribut ke i
$\boldsymbol{Y}$: Kelas yang dicari
$\boldsymbol{Y_j}$: Sub kelas Y yang dicari
$\boldsymbol{\sigma}$: Varian dari seluruh atribut
$\boldsymbol{u}$: Rata-rata dari seluruh atribut

2.8. Pengukuran Performansi

Pada tahap ini melakukan pengukuran performansi sistem *classifier* yang dibangun untuk mengetahui tingkat akurasi dengan menggunakan parameter performansi. Parameter performansi yang digunakan yaitu diantaranya nilai akurasi, *precision*, dan *recall*. Berikut adalah tabel kotingensi dari kategori Ci [14].

Tabel 1 Kotingensi untuk Prediksi dan Aktual

Kategori Ci	Kelas Prediksi		
Kelas Sebenarnya		Kelas = *Yes*	Kelas = *No*
	Kelas = *Yes*	*Tp*	*Fn*
	Kelas = *No*	*Fp*	*Tn*

Dimana:

TP (Benar Positif) : Kelas yang diprediksi *yes*, dan faktanya adalah *yes*. (hasil yang benar).
TN (Benar Negatif) : Kelas yang diprediksi *no*, dan faktanya adalah *no*. (tidak adanya hasil yang benar).
FP (Salah Positif) : Kelas yang diprediksi *yes*, dan faktanya adalah *no*. (hasil yang tidak diharapkan).
FN (Salah Negatif) : Kelas yang diprediksi *no*, dan faktanya adalah *yes*. (hasil yang meleset).

1. Akurasi

Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aslinya. Akurasi digunakan untuk mengevaluasi banyaknya label prediksi yang sesuai dengan label aktual. Semakin besar nilai akurasinya, maka performansi klasifikasi semakin baik. Berikut persamaannya [11].

$$Akurasi = \frac{TP + TN}{(TP + FP + TN + FN)} \quad (5)$$

2. Precision

Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Bila di data mining *precision* adalah jumlah dokumen yang dengan benar diklasifikasikan dalam sebuah kelas dibagi jumlah total dokumen dalam kelas tersebut. Dengan persamaan [11].

$$Precision\,(P) = \frac{TP}{(TP + FP)} \quad (6)$$

3. *Recall*

Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Dalam data mining *recall* dapat didefinisikan sebagai jumlah dokumen yang dengan benar diklasifikasikan dalam sebuah kelas dibagi jumlah total dokumen yang diklasifikasikan dalam kelas tersebut. Dengan persamaan [11].

$$Recall\,(R) = \frac{TP}{(TP + FN)} \quad (7)$$

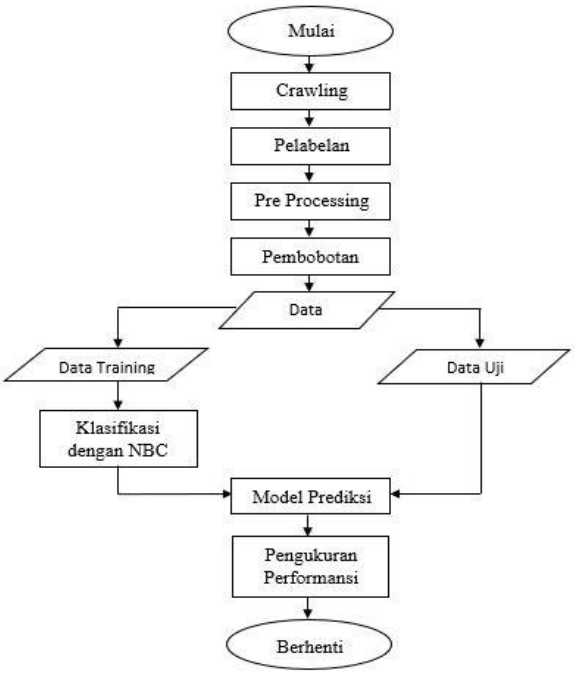
3. Perancangan Sistem

3.1. Deskripsi Umum Sistem

Deskripsi umum sistem ini menjelaskan perancangan sistem yang akan dibangun untuk melakukan penelitian tugas akhir ini tentang prediksi kepribadian *big five* dengan menggunakan metode *naïve bayes classifier*. Data yang digunakan berupa *tweet* dan informasi pengguna yang didapatkan dari API Twitter. Hasil keluaran yang diharapkan adalah model pembelajaran yang dapat digunakan untuk klasifikasi kepribadian dari Twitter.

3.2. Rancangan Sistem

Berikut gambar rancangan sitem untuk prediksi kepribadian *Big Five* pada Twitter.



Gambar 1 Rancangan Sistem Prediksi Kepribadian *Big Five* pada Twitter

3.2.1 Crawling Data

Crawling data merupakan tahap yang bertujuan untuk mendapatkan data yang akan digunakan sebagai data acuan oleh sistem yang berupa user dan tweet. Data didapatkan dengan cara mengunduh secara otomatis yang didapat dari API Twitter menggunakan implementasi bahasa pemograman PHP. Hasil keluaran dalam tahap ini yaitu data user yang sudah ada label dan data tweet.

3.2.2 Data Twitter

Pada tahap ini merupakan pengumpulan data dari hasil crawling yang diperoleh dari twitter yang selanjutnya dibagi menjadi dua data yaitu *data training* dan *data testing*.

3.2.3 Pelabelan

Pada proses ini dilakukan pemberian label kelas menjadi lima kelas kepribadian *big five* dari hasil kuesioner yang telah disebar ke koresponden dari 44 pertanyaan berdasarkan penilai BFI *(Big Five Inventory)*.

3.2.4 Preprocessing Data

Proses pada tahap *preprocessing data* ini bertujuan untuk mendapatkan data acuan yang siap diproses ke dalam sistem klasifikasi dari data tweet mentah yang diubah ke dalam bentuk yang lebih sederhana. *Preprocessing data* yang dilakukan terhadap data tweet adalah sebagai berikut.

- *Case folding*, yaitu mengubah seluruh huruf menjadi huruf kecil.
- *Tokenizing*, yaitu memisahkan setiap kata yang menyusun suatu teks menjadi satuan kata atau *term*.
- *Filtering*, yaitu pemilihan kata-kata penting menggunakan algoritma *stoplist*.
- *Stemming*, yaitu proses mengubah *term* yang berimbuhan menjadi *term* yang berbentuk kata dasar.

3.2.5 Pembobotan

Pada proses ini data yang sudah di *preprocessing data* selanjutnya akan diberi bobot dengan menggunakan metode TF-IDF dan TF-RF.

3.2.6 Klasifikasi dengan Naïve Bayes Classifier

Proses ini merupakan proses utama yang bertujuan untuk mengklasifikasikan data yang sudah melewati proses sebelumnya dengan menggunakan metode *Naïve Bayes Classifier*.

3.2.7 Model Prediksi

Proses ini merupakan sistem pembelajaran yang sudah dibuat untuk menghasilkan model prediksi kepribadian *big five*.

3.2.8 Pengukuran Performansi

Proses ini merupakan tahap akhir yaitu menghitung tingkat akrurasi dari sistem yang sudah dibuat dengan menggunakan beberapa teknik.

4. Hasil dan Analisis Penelitian

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

4.1. Data Set

Pada penelitian ini menggunakan data yang berjumlah 474.888 tweet yang telah dicrawling dengan 211 akun twitter terdiri dari lima kelas yaitu 91 akun kelas O (*Openness to Expeerience)*, kelas C (*Conscientiousness*) berjumlah 30 akun, kelas E (*Extraversion*) berjumlah 30 akun, kelas A (*Agreetableness*) berjumlah 30 akun dan kelas N (*Neuroticsm*) berjumlah 30 akun. Berikut tabel contoh data latih.

Tabel 2 Data Set

id	Akun	Follower	Following	Media_URL	URL	Mention	RT	Hashtag	Huruf_Besar	Tanda_Baca	Emoji	Kata	Rata-Rata_Kata	Karakter	Rata-Rata Karakter	label
1	AbiyogaN	16	37	19	23	99	63	13	845	31	18	141	2.311	767	12.574	O
2	Aderizkyputrii	176	1247	136	794	2856	537	51	17779	984	2569	3828	1.443	21870	8.247	A
3	Aliefirham	154	198	17	78	432	18	26	2405	181	105	743	1.351	4138	7.524	E
4	Annrahma	250	244	136	169	2290	298	129	13178	1225	1313	3153	1.098	16879	5.877	C

…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
211	ayu_sri ayuni	713	279	16	250	469	95	108	2997	163	276	282	0.48	1416	2.412	N

4.2. Hasil Uji

Pada penelitian membagi percobaan dari perilaku sosial berdasarkan data yang digunakan periodik yaitu membagi data menjadi beberapa kategori dengan cara diambil dari kuartil, desil, dan persentil, percobaan menggunakan data yang dirubah menjadi 2 kategori, 3 kategori, 5 kategori, dan gabungan dari 2, 3, dan 5 kategori dilakukan sebanyak 5 kali dan diambil nilai tertinggi dari masing-masing kategori data.

Tabel 3 Hasil akurasi data

No	Data set	Kategori	Akurasi
1	90:10	2	34,92%
2	80:20	2	39,28%
3	70:30	2	41,49%
4	60:40	2	43,65%
5	90:10	3	33,33%
6	80:20	3	36,3%
7	70:30	3	42,17%
8	60:40	3	44,44%
9	90:10	5	34,92%
10	80:20	5	38,09%
11	70:30	5	35,37%
12	60:40	5	40,47%
13	90:10	Gabungan	35,97%
14	80:20	Gabungan	38,69%
15	70:30	Gabungan	45,57%
16	60:40	Gabungan	**46,93%**

Dari percobaan berdasarkan perilaku sosial nilai akurasi tertinggi terdapat pada data gabungan kategori yaitu 46,93% dengan rasio data latih dan data uji 60:40, selanjutnya data tersebut diuji coba beberapa skenario dengan merubah atribut yang digunakan untuk melihat pengaruh dari atribut terhadap nilai akurasi.

Tabel 4 Hasil Akurasi dari banyak skenario pada Pendekatan Perilaku Sosial

No	Atribut															Akurasi
	Jumlah Tweet	Follower	Following	Media URL	URL	Mention	RT	Hashtag	Kata	Tanda Baca	Emoji	Rata-Rat Karakter	Huruf Besar	Karakter	Rata-Rata Karakter	
1	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya								38,09%
2	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya							38,88%
3	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya						42,85%
4	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya					46,03%
5	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya				46,03%
6	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya			44,44%
7	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya		44,44%
8	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	Ya	**49,20%**

Percobaan pada **Tabel 4** ini menggunakan perilaku sosial dari gabungan kategori data dan 15 atribut yaitu jumlah *tweet*, *follower*, *following*, media url, *url*, *mention*, *rt*, *hashtag*, huruf besar, tanda baca, emoji, kata, karakter, rata-rata karakter. Dari percobaan diatas nilai akurasi tertinggi didapat dari percobaan menggunakan semua atribut dengan nilai akurasi sebesar 49.20%.

Dari 145 akun yang telah dianalisa dengan menggunakan LIWC didapat 31 akun sesuai prediksi dengan kelas hasil survei. Kelas paling banyak berdasarkan LIWC yaitu *agreetableness* dan *neuroticsm* karena nilai korelasi LIWC paling tinggi kelas tersebut.

Tabel 5 Hasil Analisa Menggunakan LIWC

No	Nama	*Openness*	*Conscientiousness*	Extraversion	Agreeableness	Neuroticism	Kelas LIWC	Kelas Survei
1	adityarestuprat	95	30	70	48	53	Openness	Openness
2	amandaanio	137	79	221	432	153	Agreetableness	Agreetableness

3	ame_rahmi	187	298	482	794	278	Agreetabl eness	Agreetabl eness
4	vinzz_julian1	55	167	321	548	73	Extraversi on	Extraversi on
...	...	...	...	...	...	...	...	...
145	anitarhmlia	231	172	219	378	231	Openness	Agreetabl eness

Dari percobaan menggunakan LIWC dari 145 akun twitter dihasilkan akurasi 20,91% dengan jumlah kelas yang sesuai antara kelas hasil LIWC dengan kelas hasil survei yaitu 31 akun yang terdiri dari 2 kelas *openness*, 3 kelas *extraversion*, 19 kelas *Agreetableness*, dan 7 kelas n*euroticism.*

Pada percobaan ini dilakukan *preprocessing* dan pembobotan dengan perhitungan TF-IDF *(Term Frequency-Inverse Document Frequency)* dan TF-RF *(Term Frequency-Relevance Frequency)*. Dimana kata pada setiap *tweet* di akun 1, 2, 3…N akan dihitung berapa banyak kemunculannya (TF N). Selanjutnya nilai dari TF 1, 2, 3..N akan dikalikan dengan hasil perhitungan RF (*Frequency-Relevance Frequency*) dan IDF *(Inverse Document Frequency)*. Nilai dari hasil pengalian pada akun TF-RF 1, 2 ,3… N dan TF-IDF 1, 2, 3…N menghasilkan bobot untuk setiap kata.

Tabel 6 Contoh Kata Hasil Pembobotan TF-RF dan TF-IDF

Akun	Kata							
	mantap	**kuasa**	**guna**	**kere**	**keluar**	**cari**	**nyata**	**luar**
TF 1	2	0	4	0	2	6	4	0
TF 2	2	0	1	0	2	2	7	4
TF 3	0	0	4	0	5	11	9	4
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**
TF N	0	0	0	0	0	1	0	0
RF	6.1873	6.1429	6.1790	6.1349	6.1779	6.1879	6.1792	6.1345
IDF	2.989	4.0134	2.5134	4.2344	3.1234	4.134	4.1976	3.1349
TF_RF 1	12.226	0	24.7238	0	12.1234	38.7184	27.7334	0
TF_RF 2	12.226	0	6.1927	0	12.3558	12.2398	43.2824	24.5844
TF_RF 3	0	0	24.7988	0	30.9795	64.3234	58.6234	24.5834
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**
TF_RF N	0	0	0	0	0	6.1981	0	0
TF_IDF 1	5.889	0	10.2348	0	7.99	22.3234	16.6168	0
TF_IDF 2	5.8552	0	2.5467	0	6.92	8.1166	29.1459	13.2348
TF_IDF 3	0	0	10.1833	0	15.3	44.6234	32.4234	13.897
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**	**…**
TF_IDF N	0	0	0	0	0	4.0523	0	0

Tabel 7 dan 8 Hasil akurasi pendekatan TF-IDF dan TF-RF

No	Data set	Jumlah Kata	Akurasi
1	60:40	10	42,85%
2	60:40	20	46,82%
3	60:40	30	**50,79%**
4	70:30	10	42,85%
5	70:30	20	42,17%
6	70:30	30	45,57%
7	80:20	10	39,28%
8	80:20	20	32,73%
9	80:20	30	37,5%
10	90:10	10	34,92%
11	90:10	20	41,79%
12	90:10	30	43,91%

No	Data set	Jumlah Kata	Akurasi
1	60:40	10	42,85%
2	60:40	20	46,82%
3	60:40	30	**50,79%**
4	70:30	10	42,85%
5	70:30	20	42,17%
6	70:30	30	45,57%
7	80:20	10	38,77%
8	80:20	20	32,73%
9	80:20	30	38,09%
10	90:10	10	34,92%
11	90:10	20	41,79%
12	90:10	30	43,91%

Dari **Tabel 7** dan **Tabel 8** percobaan melalui pendekatan linguistik dengan TF-IDF didapatkan hasil akurasi terbaik di nilai 50,79 % dengan rasio data latih dan data uji 60:40 dan menggunakan 30 kata yang sering muncul, sedangkan dengan TF-RF didapatkan hasil terbaik yang sama di nilai 50,79 % dengan rasio data latih dan data uji 60:40 dan menggunakan 30 kata yang sering muncul.

Selanjutnya percobaan dengan menggabungkan dari pendekatan perilaku sosial dengan pendekatan linguistik dan menggunakan data set gabungan kategori dengan rasio data latih dan data uji 60:40 serta menggunakan 30 kata yang sering muncul.

Tabel 9 Hasil Akurasi dari perpaduan pendekatan

No	Data Set	Perilaku Sosial	TF-IDF	TF-RF	Akurasi
1	60:40	Ya	Ya		53,96%
2	60:40	Ya		Ya	53,96%
3	60:40	Ya	Ya	Ya	52,38%
4	60:40		Ya	Ya	40,47%

Dari **Tabel 9** percobaan dengan perpaduan pendekatan diperoleh nilai akurasi terbaik pada 53,96% dari percobaan perpaduan pendekatan sosial dengan TF-IDF dan perpaduan pendekatan sosial dengan TF-RF.

4.3. Analisis Penelitian

Dari hasil percobaan berdasarkan pendekatan perilaku sosial yang terdiri dari beberapa kategori dan 15 atribut didapatkan nilai akurasi tertinggi didapat dari percobaan menggunakan gabungan kategori dengan nilai yaitu 46,93% dengan rasio data latih dan data uji 60:40. Percobaan dengan merubah beberapa atribut diperoleh nilai akurasi terbaik sebesar 49.20% dengan menggunakan semua atribut.

Sedangkan percobaan menggunakan LIWC didapat 31 akun sesuai prediksi dari 145 akun dengan kelas hasil survei. Kelas paling banyak berdasarkan LIWC yaitu *agreetableness* dan *neuroticsm* karena nilai korelasi LIWC paling tinggi kelas tersebut. Pada pecobaan dengan pendekatan linguistik didapat nilai akurasi terbaik dari pendekatan TF-IDF dan TF-RF dengan nilai 50,79% dengan rasio data latih dan data uji 60:40. Kemudian untuk perpaduan dua pendekatan perilaku sosial dan linguistik didapatkan nilai akurasi terbaik 53,96% menggunakan data set dengan rasio data uji dan data latih 60:40 dengan gabungan kategori data dan menggunakn salah satu TF-IDF atau TF-RF serta dengan 30 kata yang paling sering mucul.

5. Kesimpulan dan Saran

Dari hasil penelitian yang dilakukan ini dapat disimpulkan penelitian karakter pengguna Twitter dapat diprediksi melalui fitur-fitur yang ada pada Twitter walaupun tidak semua prediksi sesuai dengan kepribadian *Big Five*, hal ini dapat disebabkan jumlah data pada survei banyak kelas *Openness to Experience* sehingga model prediksi cenderung memprediksi kelas *Openness to Experience*. Sedangkan untuk LIWC dapat memprediksi namun cenderung memprediksi kelas *agreetableness* dan *neuroticsm*, dan masih memerlukan penelitian yang terkait dengan bahasa. Percobaan dengan pendekatan linguistik dengan pembobotan TF-IDF hampir sama nilai akurasinya dengan pembobotan TF-RF, dan percobaan dengan perpaduan pendekatan perilaku sosial dan pendeketan linguistik meningkatkan nilai akurasi dengan menggunakn salah satu saja pembobotan antara TF-IDF atau TF-RF.

Saran untuk penelitian selanjutnya disarankan penggunaan data yang seimbang antar kelas dan sesuai dengan kepribadian, menambahkan kamus kata pada LIWC dan mengunnakan metode kepribadian selain *big five*. Melihat dari percobaan jika data yang digunakan berupa data numerik maka hasil yang didapat dari klasifikasi Naive bayes classifier rendah, sedangkan data yang berupa kategori nilai akurasi meningkat karena klasifikasi Naive bayes classifier paling baik digunakan untuk data yang berupa kategori. Dan untuk LIWC masih diperlukan penelitian untuk nilai yang digunakan sesuai dengan Bahasa Indonesia.

6. Daftar Pustaka

[1] G. Saucier and L. R. Goldberg, "What is beyond the big five?," *J. Pers.*, vol. 66, no. 4, pp. 495–524, 1998.

[2] Y. Liu, J. Wang, and Y. Jiang, "PT-LDA: A latent variable model to predict personality traits of social network users," *Neurocomputing*, vol. 210, pp. 155–163, 2016.

[3] J. Golbeck, "Predicting Personality with Social Media," *Hum. Factors*, pp. 253–262, 2011.

[4] 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.

[5] A. T. Damanik and K. Kunci, "Prediksi Kepribadian Big 5 Pen gguna Twitter dengan Support Vector Regression," no. 1981, pp. 14–22, 1999.

[6] "Perkembangan Media Sosial di Indonesia - PakarKomunikasi.com." [Online]. Available: https://pakarkomunikasi.com/perkembangan-media-sosial-di-indonesia. [Accessed: 26-Feb-2018].

[7] "Algoritma Naive Bayes" [Online]. Available: https://informatikalogi.com/algoritma-naive-bayes/. [Accessed: 26-Mei-2018].

[8] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from twitter," *Proc. - 2011 IEEE Int. Conf. Privacy, Secur. Risk Trust IEEE Int. Conf. Soc. Comput. PASSAT/SocialCom 2011*, pp. 149–156, 2011.

[9] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," *Proc. - 2011 IEEE Int. Conf. Privacy, Secur. Risk Trust IEEE Int. Conf. Soc. Comput. PASSAT/SocialCom 2011*, pp. 180–185, 2011.

[10] "Teori Kepribadian Model Lima Besar (Big Five Personality) - IPQI." [Online]. Available: https://www.ipqi.org/teori-kepribadian-model-lima-besar-big-five-personality/. [Accessed: 24-Feb-2018].

[11] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data*. 2014.

[12] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.

[13] A. Manuscript and E. Dysfunction, "NIH Public Access," vol. 25, no. 8, pp. 713–724, 2015.

[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[15] Wu Haibing dan Gu Xiaodong, *Reducing Over-Weighting in Supervised Term Weighting for Sentiment Analysis*.2014.