Prediksi *Big Five Personality* dengan *Term Frequency Inverse Document Frequency* (TF – IDF) Menggunakan Metode *Logistic Regression* pada Pengguna Twitter

Rendo Zenico[1], Erwin Budi Setiawan[2], Fida Nurmala Nugraha[3]

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

Abstrak

Kepribadian atau *personality* bisa didefinisikan sebagai keseluruhan cara seseorang bereaksi dan berinteraksi dengan lingkungan maupun dengan individu lainnya. Penelitian yang berkaitan dengan kepribadian seseorang sudah banyak dilakukan para ahli untuk kepentingan tertentu. Penelitian terhadap kepribadian seseorang dalam penggunaan media sosial juga sudah mulai banyak dikembangkan. Salah satu media sosial yang digunakan untuk mengamati dan meneliti kepribadian dari seseorang adalah Twitter. Dengan banyaknya pengguna aktif pada Twitter, setiap individu pasti memiliki ciri yang berbeda dalam menggunakan akun Twitter mereka. Pada penelitian tugas akhir ini, penulis membangun sistem klasifikasi kepribadian pengguna Twitter. Penelitian yang dilakukan menggunakan dua pendekatan, yaitu pendekatan linguistik dan pendekatan perilaku sosial dengan menggunakan fitur dari Twitter itu sendiri. Data yang digunakan adalah data dari 143 pengguna Twitter dengan jumlah 351,197 *tweet* dengan rasio perbandingan data latih dan data uji 70%:30%. Menggunakan pembobotan *Term Frequency Inverse Document Frequency* (TF – IDF) dan *Logistic Regression* sebagai algoritma klasifikasi, akurasi yang dihasilkan oleh sistem yang dibangun pada tugas akhir ini 69% untuk pendekatan perilaku sosial dan 76,20% untuk pendekatan linguistik dan pendekatan perilaku sosial.

Kata kunci : twitter, pembobotan, TF – IDF, *logistic regression*

Abstract

Personality can be defined as the overall way in which someone reacts and interacts with the environment and with other individuals. Research related to someone personality has been carried out by experts for certain purposes. Research on someone personality in the use of social media has also begun to be developed. One of the social media used to observe and examine someone personality is Twitter. With many active users on Twitter, each individual must have different characteristics in using their Twitter account. In this thesis research, the authors built a personality classification system for Twitter users. Research carried out using two approaches, namely the linguistic approach and features of Twitter itself. The data used are data from 143 Twitter users with a total of 351,197 tweets with a comparison ratio of training data and test data 70%:30%. Using weighting Term Frequency Inverse Document Frequency (TF - IDF) and Logistic Regression as classification algorithms, the accuracy produced by the system built in this final project reaches 69% for the social behavior approach and 76,20% for the linguistic approach and social behavior approach.

Keywords: twitter, weighting, TF – IDF, logistic regression

1. Pendahuluan

Salah satu teori yang menjelaskan tentang kepribadian seseorang adalah "*Big Five Personality Traits Model*" atau "Model Lima Besar Sifat Kepribadian" yang dikemukakan oleh Lewis Goldberg. Menurut Goldberg, ada lima dimensi kunci yang bisa mencirikan kepribadian seseorang. Lima dimensi tersebut yaitu *Openness to Experience*, *Conscientiousness*, *Extraversion*, *Agreeableness*, dan *Neuroticism*. Banyak penelitian yang dikembangkan terkait dengan prediksi kepribadian pengguna Twitter. Pada penelitian [2] prediksi kepribadian pengguna Twitter dilakukan menggunakan metode *Support Vector Regression*. Penelitian [3] prediksi kepribadian yang dilakukan menggunakan metode Regresi.

Pada penelitian tugas akhir ini, penulis menggunakan metode *Logistic Regression* sebagai algoritma untuk melakukan prediksi klasifikasi kepribadian pengguna Twitter. Metode ini dikombinasikan dengan TF – IDF untuk menghitung bobot, yang digunakan dalam pendekatan fitur linguistik. Selain itu, pada penelitian tugas akhir ini penulis juga menggunakan pendekatan terhadap fitur dari Twitter, dengan melihat fitur – fitur apa saja dari Twitter yang bisa mempengaruhi hasil klasifikasi terhadap kepribadian penggunanya.

2. Studi Terkait

2.1 *Big Five Personality*

Banyak definisi yang dikemukakan oleh para ahli tentang kepribadian manusia seperti Gordon Allport, Adolf Heuken, maupun Krech dan Crutchfield. Pada dasarnya, kepribadian bisa didefinisikan sebagai cara keseluruhan seseorang bereaksi dan berinteraksi dengan lingkungan maupun individu lainnya. Faktor yang

mempengaruhi sifat kepribadian seseorang bisa diakibatkan karena faktor genetis keturunan dan faktor lingkungan. Faktor lingkungan yang mempengaruhi bisa diakibatkan karena norma yang berlaku maupun cara seseorang tersebut dibesarkan atau bertumbuh pada lingkungannya.

Sifat adalah bagian dari kepribadian. Bisa diartikan sebagai dimensi dari kepribadian yang terkait dengan respons atau reaksi seseorang yang relatif konsisten dalam hal menyesuaikan dirinya. Lewis Goldberg memperkenalkan teori sifat kepribadian "*Big Five Personality Traits Model*" atau "Model Lima Besar Sifat Kepribadian". Goldberg menjelaskan ada lima dimensi kunci kepribadian. Lima dimensi kepribadian ini memiliki cirinya masing – masing yang menggambarkan karakteristik individunya.

1. *Openness to Experience* (Terbuka terhadap hal – hal baru).
   Individu ini memiliki ketertarikan terhadap hal – hal baru. Keinginan untuk mengetahui dan mempelajari hal yang baru. Individu dengan kepribadian ini biasanya lebih kreatif, imajinatif, penasaran, intelektual dan mempunyai pemikiran yang luas.
2. *Conscientiousness* (Sifat berhati – hati).
   Individu dengan kepribadian ini cenderung penuh pertimbangan ketika akan mengambil keputusan. Karena penuh pertimbangan, individu dengan kepribadian ini biasanya bisa dipercaya, dapat diandalkan, bertanggung jawab, tekun dan mempunyai orientasi pada target yang ingin dicapai.
3. *Extraversion* (Ekstarversi).
   Dimensi kepribadian ini mempunyai hubungan dengan tingkat kenyamanan dalam berinteraksi dengan orang lain. Karakteristiknya adalah cenderung mudah bergaul, bersosialisasi, suka hidup berkelompok dan cenderung lebih tegas.
4. *Agreeableness* (Mudah bersepakat).
   Individu dengan kepribadian ini cenderung untuk menghindari konflik dan lebih mudah patuh dengan individu lainnya. Sisi positif dari individu dengan karakteristik ini adalah dapat bekerja sama, lebih bersifat baik, cenderung hangat, penuh kepercayaan dan suka membantu.
5. *Neuroticism* (Neurotisme).
   Individu dengan karakteristik ini cenderung lebih mampu menahan tekanan. Individu dengan dimensi ini lebih memiliki stabilitas emosional yang baik, lebih tenang menghadapi masalah, pendirian teguh dan lebih percaya diri

Teori sifat kepribadian individu ini sudah banyak dimanfaatkan dalam berbagai bidang, diantaranya adalah untuk melihat evaluasi kinerja anggota dewan [4] maupun sebagai pedoman untuk penentuan karir bagi remaja usia sekolah [5].

2.2 Pendekatan Perilaku Sosial

Pendekatan ini mendefinisikan kepribadian berdasarkan tingkat keaktifan pengguna Twitter dalam menggunakan akun Twitter mereka. Penelitian [6] menjelaskan fitur yang mempengaruhi tingkat perilaku sosial pengguna Twitter.

A.	*Follower*	:	Pengguna lain yang mengikuti akun pengguna yang diacu.
B.	*Following*	:	Pengguna yang diacu menjadi *follower* pengguna lain.
C.	Jumlah *Mention*	:	Tingkat interaksi pengguna Twitter dengan pengguna lain.
D.	Jumlah *Hashtag*	:	Keterlibatan pengguna terhadap topik tertentu, adanya penggunaan karakter '#'.
E.	Jumlah *Reply*	:	*Mention* dari pengguna lain kepada pengguna Twitter yang diacu.
F.	Jumlah URL	:	Tautan berupa informasi alamat *link* yang dibagikan pengguna.
G.	Jumlah Kata dalam *Tweet*	:	Jumlah kata dalam *tweet*, total kata yang menyusun *tweet*.

Selain fitur di atas, terdapat juga fitur lain yang menjadi bahan pertimbangan. Fitur ini juga akan dianalisis untuk menunjukkan pengaruh tingkat keaktifan perilaku sosial pengguna Twitter.

H.	Jumlah *Retweet*	:	Jumlah unggahan kembali *tweet* dari pengguna lain.
I.	Jumlah Media URL	:	Jumlah tautan berupa gambar, video maupun media.
J.	Jumlah Tanda Baca	:	Tanda baca yang digunakan, yang dihitung hanya '!' dan '?'.
K.	Jumlah Emoji	:	Karakter unik yang bisa menggambarkan emosi pengguna ketika menggunggah *tweet*.
L.	Rata – Rata Kata	:	Rata – rata kata dari *tweet* pengguna yang sudah di*crawl*. *Retweet* tidak dihitung
M.	Jumlah Huruf Besar	:	Huruf kapital yang digunakan saat menulis *tweet*.
N.	Jumlah Karakter	:	Susunan huruf, simbol - simbol yang menyusun sebuah *tweet*.
O.	Rata – Rata Karakter	:	Rata – rata karakter yang dituliskan pengguna di *tweet* yang sudah dicrawling. *Retweet* tidak dihitung.

2.3 Pendekatan Linguistik

Pendekatan linguistik adalah pendekatan fitur kata dari *tweet*. Pendekatan linguistik melihat bagaimana hubungan antara kata dari hasil kumpulan *tweet* terkait dengan kepribadian pengguna Twitter. Pendekatan

linguistik dilakukan dengan cara mengurai *tweet* menjadi *term* kata *unigram*. *Term* kata *unigram* ini kemudian dihitung bobotnya dengan TF – IDF.

2.4 *Prepocessing*

Preprocessing merupakan proses mengubah data menjadi terstruktur sehingga bisa digunakan sesuai dengan kebutuhan. Pada penelitian tugas akhir ini, semua *tweet* yang diambil dari pengguna Twitter akan di*preprocessing* dengan 4 tahapan :

a. *Case Folding* : Merubah semua huruf menjadi non – kapital.
b. *Tokenizing* : Mengurai kalimat menjadi kata, menghilangkan tanda baca, spasi serta karakter yang tidak perlu. Pada penelitian tugas akhir ini digunakan pemisahan kata *unigram*.
c. *Filtering* : Memilih kata penting dari hasil *tokenizing* dengan memanfaatkan *stoplist* atau kamus yang sudah didefinisakan.
d. *Stemming* : Membentuk kata menjadi bentuk kata dasarnya.

Tabel 1. Ilustrasi *Preprocessing*

KALIMAT	*CASE FOLDING*	*TOKENIZING*	*FILTERING*	*STEMMING*
Hari ini pasti meraih KEMENANGAN!!!	hari ini pasti meraih kemenangan!!!	hari | ini | pasti | meraih | kemenangan	hari | pasti | meraih | kemenangan	hari | pasti | raih | menang

2.5 *Term Weighting* (Pembobotan)

Pembobotan merupakan pemberian nilai atau bobot untuk *term* atau kata pada suatu dokumen. Nilai bobot *term* akan menjadi penentu klasifikasi suatu teks. Pembobotan yang digunakan pada penelitian tugas akhir ini menggunakan TF – IDF dan dilakukan sebelum menjalankan proses klasifikasi suatu teks.

2.5.1 Term Frequency Inverse Document Frequency (TF – IDF)

Salton mengusulkan metode ini menggabungkan metode *Term Frequency* (TF) dan *Inverse Document Frequency* (IDF). TF adalah bobot dari suatu kata (t) dalam suatu dokumen (d). Pendekatan paling sederhana dari konsep ini dengan menyatakan bobot suatu kata (t) dengan jumlah kemunculannya pada dokumen (d).

$$tf_{t,d} = \begin{cases} \log f_{t,d}\,, & Jika\ f_{t,d} > 0 \\ 0, & Jika\ kata\ tidak\ muncul \end{cases} \quad (1)$$

Keterangan:
$tf_{t,d} = term\ frequency$
$f_{t,d} = jumlah\ kemunculan\ kata\ (t)\ dalam\ dokumen\ (d)$

TF menilai suatu dokumen sebagai *bag of words* (kantong kata). Urutan dari kemunculan kata diabaikan dan hanya menilai jumlah kemunculan dari kata itu saja yang dianggap penting. Konsep ini memiliki kelemahan karena semua kata dianggap setara, dan mengakibatkan relevansi kata menjadi tinggi jika muncul dibanyak dokumen. Tingginya frekeunsi kemunculan suatu kata tidak selalu menyatakan bahwa kata tersebut penting.

IDF dibuat untuk mengurangi efek dari kata yang frekuensinya tinggi dalam sekumpulan dokumen. Dasar idenya adalah menurunkan bobot dari frekuensi total kemunculan kata disemua dokumen yang tinggi.Semakin banyak kata tersebut muncul pada sekumpulan dokumen, maka semakin rendah bobotnya.

$$idf_t = \log \frac{N}{N_t} \quad (2)$$

Keterangan:
$idf_t = inverse\ document\ frequency$
$N = jumlah\ keseluruhan\ dokumen$
$N_t = jumlah\ dokumen\ yang\ memuat\ kata\ (t)$

Algoritma TF – IDF merupakan algoritma *information retrieval*. Algoritma ini digunakan untuk menentukan bobot dari kata (t) pada suatu dokumen (d).

$$tfidf_{t,d} = tf_{t,d} \times idf_t \quad (3)$$

Keterangan:
$tfidf_{t,d} = bobot\ term\ atau\ kata$
$tf_{t,d} = term\ frequency\ kata\ (t) pada\ dokumen\ (d)$
$idf_t = inverse\ document\ frequency\ kata\ (t)$

2.6 *Logistic Regression*

Logistic Regression merupakan pendekatan untuk membuat model prediksi. *Logistic Regression* merupakan salah satu *Supervised Classification Algorithm* yang berangkat dari algoritma *Linear Regression*. Persamaan *Logistic Regression* dapat dilihat pada persamaan (4):

$$P(Y|X) = \frac{\exp[\beta_0 + \sum_i \beta_i X_i]}{1 + \exp[\beta_0 + \sum_i \beta_i X_i]} \quad (4)$$

Keterangan:
P(Y|X) = *Peluang kelas Y pada observasi X*
β0 = *Bias*
βi = *Vektor bobot*
Xi = *Variabel pengamatan*

Logistic Regression dapat dibedakan menjadi dua metode. Pertama *Binary Logistic Regression* digunakan untuk dua kemungkinan variabel Y. *Multinomial Logistic Regression* digunakan untuk lebih dari dua kemungkinan variabel Y. Probabilistik *Multinomial Logistic Regression* menurut [7] :

$$P(c|x) = \frac{exp(\sum_{i=1}^{N} k_i f_i(c,x))}{\sum_{c\prime \in C} exp(\sum_{i=1}^{N} k_i f_i(c\prime,x))} \quad (5)$$

Keterangan:
P(c|x) = *Peluang kelas c pada observasi x*
k_i = *Vektor bobot*
$f_i(c,x)$ = *fitur i untuk kelas tertentu c pada observasi x*

2.7 Performansi Sistem

Pengukuran performansi sistem pada penelitian tugas akhir ini menggunakan *precision, recall* dan akurasi berdasarkan *Confussion Matrix* untuk skenario *multiclass*. *Confussion Matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi sebenarnya.

Tabel 2. *Confussion Matrix* untuk Skenario *Multiclass*

	Predicted: NO	***Predicted: YES***
Actual: NO	TN_i	FP_i
Actual: YES	FN_i	TP_i

Keterangan:
- TP_i adalah *True Positive*, jumlah data positif yang terklasifikasi dengan benar oleh sistem untuk kelas ke – i.
- TN_i adalah *True Negative*, jumlah data negatif yang terklasifikasi dengan benar oleh sistem untuk kelas ke – i.
- FN adalah *False Negative*, jumlah data negatif namun terklasifikasi salah oleh sistem untuk kelas ke – i.
- FP adalah *False Positive*, jumlah data positif namun terklasifikasi salah oleh sistem untuk kelas ke – i.
- *I* adalah jumlah kelas.

2.7.1 Precision

Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem, didefinisikan:

$$Precision = \frac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} (FP_i + TP_i)} * 100\% \quad (6)$$

2.7.2 Recall

Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi, didefinisikan:

$$Recall = \frac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} (TP_i + FN_i)} * 100\% \quad (7)$$

2.7.3 Akurasi

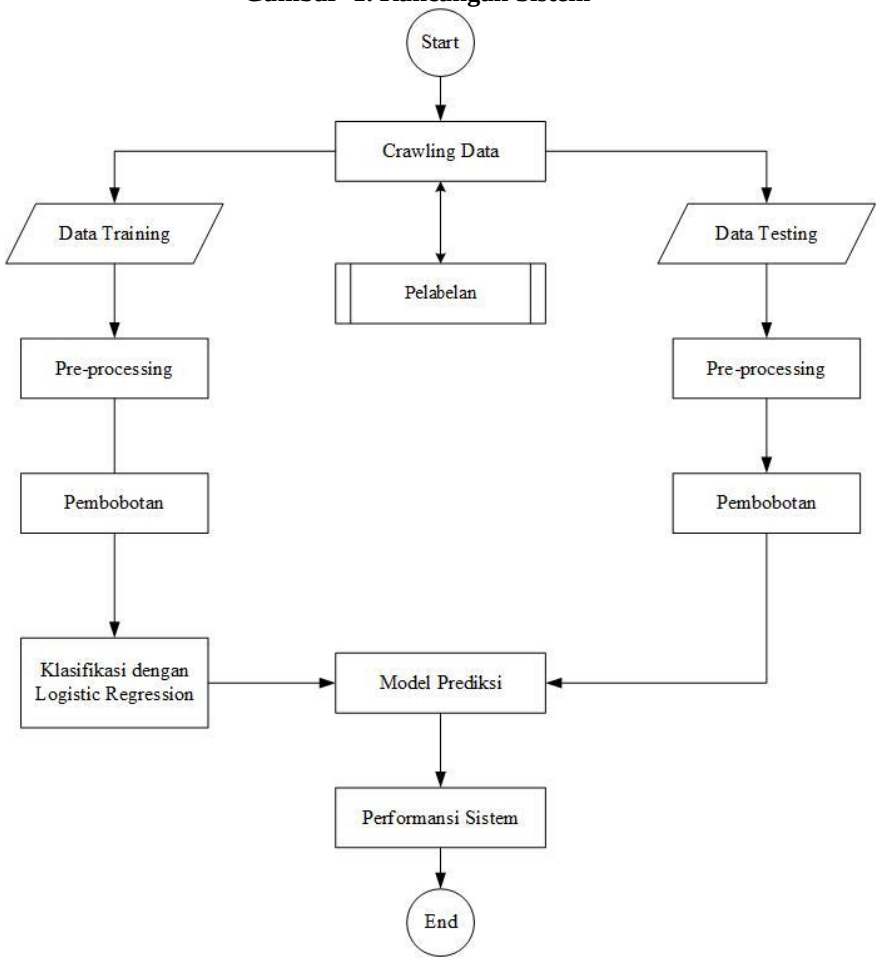
Akurasi adalah sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual, didefinisikan:

$$Akurasi = \frac{\sum_{i=1}^{1} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{I} * 100\% \quad (8)$$

3. Sistem yang Dibangun

Rancangan sistem yang dibangun dalam penelitian tugas akhir ini bisa dilihat pada **Gambar 1.**

Gambar 1. Rancangan Sistem



1. *Crawling* Data
   Crawling data merupakan proses pengambilan data akun pengguna Twitter. Data yang diambil berupa *tweet*, *follower*, *following*, jumlah *mention*, jumlah *hashtag*, jumlah *reply*, jumlah URL, jumlah kata dalam *tweet*, jumlah *retweet*, jumlah media URL, jumlah tanda baca, jumlah emoji, rata-rata kata, jumlah huruf besar, jumlah karakter, rata – rata karakter. Untuk *tweet*, jumlah *tweet* maksimal yang diambil dari setiap pengguna adalah 3200 *tweet* terbaru. Untuk akun yang akan di *crawling* berasal dari penguna Twitter yang sudah mengisi kuisioner yang telah disebar. Dari 220 responden, akun pengguna Twitter yang bisa di*crawl* hanya berjumlah 143 akun.
2. Pelabelan
   Pelabelan terhadap akun pengguna Twitter yang sudah di crawl dilakukan berdasarkan hasil kuisioner yang sudah di isi responden. Kuisioner yang dibagikan berisi 44 pertanyaan kepribadian *Big Five Inventory* yang sudah di terjemahkan ke Bahasa Indonesia [8], [9].
3. Distribusi Data
   Distribusi data merupakan pembagian komposisi data sebagai data latih dan data uji untuk dicoba dalam pengujian model prediksi yang dibangun. Pada penelitian tugas akhir ini, penulis membagi komposisi data latih dan data uji 70%:30%.
4. *Preprocessing*
   Preprocessing bertujuan mengubah data yang tidak terstruktur menjadi sesuai struktur yang diinginkan agar dapat diolah sesuai kebutuhan. Tahapan ini mencakup *Case Folding, Tokenizing, Filtering* dan *Stemming*. Selain itu, *retweet* dan URL juga dihapuskan karena kurang dapat menjelaskan sifat kepribadian akun pengguna Twitter tersebut.
5. Pembobotan
   Tweet yang sudah di*preprocessing* akan diberikan nilai atau bobot menggunakan pembobotan TF – IDF. Tujuannya adalah untuk mengukur pengaruh kata dari dokumen *tweet* terhadap klasifikasi kepribadian pengguna Twitter.

6. Klasifikasi dengan *Logistic Regression*
   Data yang sudah di*preprocessing* akan masuk untuk pengujian klasifikasi. Pada proses ini, data latih akan dimasukkan kedalam perhitungan *Logistic Regression*. *Output* dari perhitungan ini akan menghasilkan model prediksi yang performansinya akan diuji.
7. Model Prediksi
   Model prediksi merupakan sistem klasifikasi yang sudah dibuat dengan *Logistic Regression*. Model prediksi yang sebelumnya sudah dibangun akan diuji dengan data uji yang sudah disiapkan. *Output* dari model prediksi ini akan menunjukkan nilai performansi sistem yang telah dibuat.
8. Performansi Sistem
   Perhitungan *precision*, *recall*, dan akurasi akan dilakukan pada proses ini, yang bertujuan untuk mengukur performansi sistem yang telah dibuat.

4. Evaluasi

Pada bagian ini, skenario yang dibuat akan diuji dengan data latih yang telah dibuat sebelumnya.

4.1. Dataset dan Pelabelan

Dataset yang digunakan untuk penelitian tugas akhir ini berjumlah 143 akun pengguna Twitter yang telah di*crawling*. *Tweet* yang di*crawling* berjumlah 351,197 *tweet*. Dari 143 akun yang didapat dari responden yang telah mengisi kuisioner, selanjutnya dilakukan pelabelan. Hasil pelabelan 143 data akun pengguna Twitter yang sudah di*crawling*, didapatkan jumlah label *Openness to Experience* 78 orang, *Conscientiousness* 14 orang, *Extraversion* 2 orang, *Agreeableness* 32 orang, *Neuroticsm* 17 orang. Untuk komposisi perbandingan data latih dan data uji adalah 70%:30%.

4.2. Hasil Pengujian

Hasil pengujian pada penelitian tugas akhir ini menggunakan skenario pendekatan linguistik dan perilaku sosial. Pendekatan linguistik akan menggunakan perhitungan nilai bobot dengan TF – IDF. Nilai bobot ini didapat dari jumlah kemunculan TF yang selanjutnya akan dikalikan dengan hasil perhitungan IDF. Hasil TF – IDF ini akan menghasilkan bobot untuk setiap kata.

Tabel 3. Contoh Perhitungan TF - IDF

Akun Twitter	**Dokumen Kelas (*Openness to Experience*, *Conscientiousness*, *Extraversion*, *Agreeableness*, *Neuroticsm*)**				
	Jalan	**Bisa**	**Cari**	**Dalam**	**Ingin**
TF_faidillzham	144	52	8	22	4
TFIDF_faidilzham	0,086291	0,19426	0,122721	0,337483	0,10778

Skenario pengujian dengan pendekatan fitur perilaku sosial, dengan menghitung pengaruh atribut fitur dari Twitter yang mempengaruhi hasil klasifikasi dengan *Logistic Regression* ditunjukkan **Tabel 4.**

Tabel 4. Akurasi Pendekatan Perilaku Sosial Menggunakan *Logistic Regression*

Atribut	**Akurasi** %
follower, *following*, emoji	52,41
follower, *following*, emoji, media_url	54,76
follower, *following*, emoji, media_url, url	57,14
follower, *following*, emoji, media_url, url, rata2_kata	57,14
follower, *following*, emoji, media_url, rata2_kata, *hashtag*	57,14
follower, *following*, emoji, media_url, url, rata2_kata, hashtag, retweet	54,76
following, emoji, media_url, url, rata2_karakter, *retweet*, kata, rata2_kata, karakter, *mention*, tanda_baca, huruf_besar, *hashtag*	64.36
follower, following*, media_url, url, rata2_karakter, *retweet*, kata, rata2_kata, karakter, *mention*, tanda_baca, huruf_besar, *hashtag	**69**

Pada skenario pendekatan perilaku sosial, akurasi yang didapatkan 69%, dengan menggunakan semua fitur atribut dari Twitter terkecuali emoji.

Untuk skenario TF – IDF dengan pendekatan perilaku sosial, akurasi yang didapat lebih baik dibandingkan dengan pendekatan perilaku sosial. Pendekatan linguistik dengan TF – IDF dan pendekatan perilaku sosial menghasilkan akurasi 76,20%.

Tabel 5. TF - IDF dengan Pendekatan Perilaku Sosial

Atribut	**Akurasi** %
TF – IDF, *follower*, *following*, media_url, url, rata2_karakter, *retweet*, kata, rata2_kata, karakter, *mention*, tanda_baca, huruf_besar, *hashtag*	76,20

4.3 Analisis Hasil Pengujian

Dari skenario yang sudah dilakukan pada penelitian tugas akhir ini menghasilkan akurasi yang berbeda. Pada

pengujian pendekatan perilaku sosial dengan atribut Twitter, akurasi yang terbaik yang dihasilkan 69%. Fitur atribut yang berpengaruh *follower*, *following*, media_url, url, rata2_karakter, *retweet*, kata, rata2_kata, karakter, *mention*, tanda_baca, huruf_besar, *hashtag*.

Pengujian pendekatan linguistik dengan TF - IDF dan perilaku sosial, menghasilkan akurasi yang lebih baik. Akurasi yang didapat dari scenario ini adalah 76,20%. Penggunaan bentuk kata *unigram* pada TF – IDF juga belum mampu untuk memaksimalkan uji similiaritas data.

5. Kesimpulan

Penelitian prediksi kepribadian *Big Five Personality* menggunakan TF – IDF dan metode *Logistic Regression* menghasilkan akurasi yang berbeda. Pendekatan perilaku sosial menghasilkan akurasi sebesar 69%. Hasil ini didapat dari pemilihan atribut yang mempengaruhi yaitu *follower*, *following*, media_url, url, rata2_karakter, *retweet*, kata, rata2_kata, karakter, *mention*, tanda_baca, huruf_besar, *hashtag*.

Sedangkan pada pendekatan linguistik dan perilaku sosial, hasil akurasi yang didapat 76,20%. Hal ini menunjukkan TF – IDF mampu menaikkan akurasi prediksi untuk sistem yang dibangun. Dari total 143 data yang digunakan, untuk kelas Opennes to Experience terlalu dominan dibanding kelas lainnya, yang mempengaruhi tingkat keberagaman data. Sistem yang dibangun secara umum sudah bisa melakukan prediksi kepribadian pengguna Twitter. Akurasi yang berbeda dari skenario yang dijalankan menunjukkan adanya hal yang mempengaruhi akurasi. Sebaran data label yang tidak merata juga mengakibatkan adanya kecenderungan prediksi menuju ke kelas yang dominan.

Saran untuk penelitian selanjutnya, kuisioner untuk responder lebih merata untuk semua kelas label. Pendekatan linguistik dan perilaku sosial, secara umum menghasilkan akurasi yang baik, tetapi perlu pengembangan untuk memilih kata yang dominan disetiap kelasnya. Pendekatan perilaku sosial juga perlu dikembangkan untuk melihat pengaruh atribut lain yang bisa mempengaruhi hasil prediksi.

Daftar Pustaka

[1] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," in *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, 2011, no. October, pp. 180–185.

[2] A. T. Damanik and Masayu Leylia Khodra, "Prediksi Kepribadian Big 5 Pengguna Twitter dengan Support Vector Regression," *J. Cybermatika*, vol. 3, no. 1, pp. 14–22, 2015.

[3] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting Personality from Twitter," in *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, 2011, pp. 149–156.

[4] Hardani Widhiastuti, "Big Five Personality sebagai Prediktor Kreativitas dalam Meningkatkan Kinerja Anggota Dewan 115 Hardani," *J. Psikol.*, vol. 41, no. 1, pp. 115–133, 2014.

[5] K. Öztemel, "Career Indecisiveness of Turkish High School Students:: Associations With Personality Characteristics," *J. Career Assess.*, vol. 22, no. 4, pp. 666–681, 2014.

[6] S. Adali and J. Golbeck, "Predicting personality with social behavior," *Proc. 2012 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2012*, pp. 302–309, 2012.

[7] "Analisis Pemakaian Kemoterapi Pada Kasus Kanker Payudara Dengan Menggunakan Metode Regresi Logistik Multinomial (Studi Kasus Pasien Di Rumah Sakit 'X' Surabaya)," *J. Sains dan Seni ITS*, vol. 1, no. 1, 2012.

[8] P. J. Oliver and S. S., "The Big-Five Taxonomy: History; Measurement; and Theoretical Perspectives," *Handb. Personal. Theory Res.*, vol. 2, pp. 102–138, 1999.

[9] N. Ramadhani, "Adaptasi Bahasa dan Budaya Inventori Big Five," *J. Psikol.*, vol. 39, no. 2, pp. 189–207, 2012.