DETEKSI KEPRIBADIAN BIG FIVE PENGGUNA TWITTER DENGAN METODE C4.5

Shantika Valerin Therik[1], Erwin Budi Setiawan[2]

[1,2] Universitas Telkom, Bandung
[1]shantikavt@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

Abstrak

Twitter merupakan media sosial yang memberikan fitur bagi penggunannya untuk membaca dan menulis pesan yang disebut "*tweet*". Untuk dapat memahami kepribadian seseorang, postingan di media sosial yaitu Twitter dapat digunakan sebagai sumber informasi. Big Five dalam dunia psikologi merupakan salah satu metode untuk menginterpretasi kepribadian seseorang. Kepribadian yang sesuai memiliki dampak secara langsung pada kinerja di tempat kerja. Personality fit akan melihat bagaimana kepribadian seseorang sesuai dengan pekerjaan dan budaya perusahaan. Pada penelitian ini, metode C4.5 digunakan untuk membuat model klasifikasi kepribadian pengguna Twitter yang terdiri dari lima kelas yaitu *Openness, Consciousness, Extraversion, Agreeableness, Neuroticism.* Dataset yang diperoleh menggunakan API Twitter. Dilakukan percobaan dengan skenario mendapat rasio data dari akurasi perilaku sosial sebagai baseline, penambahan data TF-IDF dan LIWC dan menerapkan metode SMOTE yang diujikan menerapkan teknik *hyperparameter tuning* menggunakan *Grid Search* dengan perilaku sosial sebagai *baseline.* Hasil akurasi yang diperoleh dengan penambahan data TF-IDF dan LIWC sebesar 62.06% dengan kenaikan akurasi sebesar 17.24% dari *baseline* dan menggunakan metode SMOTE dapat meningkatkan nilai akurasi menjadi 76.92% dengan kenaikan nilai akurasi sebesar 32.1% dari *baseline*. Dari fitur dengan nilai akurasi terbaik dalam percobaan, dihasilkan model pohon keputusan deteksi kepribadian Big Five.

Kata kunci : Big Five, Klasifikasi C4.5, TF-IDF, LIWC

Abstract

Twitter is a social media that provides a feature for its users to read and write messages known as "tweets". To be able to understand a person's personality, posts on social media, namely Twitter, can be used as a source of information. The Big Five in psychology is one method for interpreting a person's personality. Appropriate personality has a direct impact on performance at work. Personality fit will see how a person's personality fits the job and company culture. In this study, the C4.5 method was used to create a personality classification model for Twitter users which consisted of five classes, namely Openness, Consciousness, Extraversion, Agreeableness, and Neuroticism. Dataset obtained using the Twitter API. An experiment was conducted with the scenario of getting the data ratio from the accuracy of social behavior as a baseline, adding TF-IDF and LIWC data and applying the SMOTE method which was tested using a hyperparameter tuning technique using Grid Search with social behavior as the baseline. The accuracy results obtained by adding TF-IDF and LIWC data are 62.06% with an increase in accuracy of 17.24% from the baseline and using the SMOTE method can increase the accuracy value to 76.92% with an increase in the accuracy value of 32.1% from the baseline. From the features with the best accuracy values in the experiment, the Big Five personality detection decision tree model was generated.

Keywords : Big Five, C4.5 Classification, TF-IDF, LIWC

1. Pendahuluan

Pada zaman ini, orang lebih banyak berinteraksi secara *online* melalui media sosial dibandingkan bersosialisasi secara langsung sehingga lebih banyak informasi yang terdapat pada media sosial yang mana berarti media sosial sebagai tempat bagi penggunanya untuk mengekspresikan dirinya, berkomunikasi dan bertukar informasi. Twitter merupakan media sosial yang memberikan fitur bagi penggunanya untuk membaca dan menulis pesan yang disebut "*tweet*"[3]. Informasi pada media sosial tersebut dapat digunakan untuk mendeteksi kepribadian seseorang. Kepribadian merupakan bagian kajian psikologi, pemahaman tentang perilaku, pikiran, perasaan, memakai sistematik, metode dan rasional psikologik[1].

Di era teknologi ini, untuk mengetahui kepribadian seseorang bisa dideteksi dari aktivitas percakapan pada media sosialnya karena terdapat banyak informasi tentang pengguna seperti banyak perusahaan yang menggunakan deteksi kepribadian calon karyawannya menggunakan informasi dari media sosial. Big Five dalam dunia psikologi merupakan salah satu metode untuk menginterpretasikan kepribadian seseorang [2]. Big Five sebagai acuan dalam mendeteksi kepribadian pengguna Twitter dimana data yang terdapat dalam media sosial Twitter berbentuk teks. Kepribadian Big Five terdiri dari *Openness*, *Consciousness*, *Extraversion*, *Agreeableness*, *Neuroticism* (OCEAN)[4].

Peneliti sebelumnya menggunakan pola tanda tangan dalam deteksi kepribadian yang dilakukan oleh Bhakti Prasetya Utama. Metode yang digunakan adalah *Support Vector Machine* (SVM) dengan hasil rata – rata akurasi 60,7% dan dengan metode Rule Based diperoleh hasil rata – rata akurasi sebesar 52,8%[5]. Penelitian dengan metode

yang sama yaitu *Support Vector Machine* (SVM) dan metode *Principal Component Analysis* (PCA) dilakukan oleh Maulana Amsor diperoleh hasil rata – rata akurasi sebesar 70,9%[6].

C4.5 merupakan algoritma dalam *data mining* yang penerapannya akan menghasilkan pohon keputusan (*decision tree*). Pohon keputusan tersebut terdiri atas aturan-aturan yang bercabang atau bertingkat berdasarkan pada data dan hasil kalkulasi *entropy* dan *information gain*[7].

Pada penelitian ini, telah dilakukan klasifikasi menggunakan algoritma C4.5 untuk mendapatkan model deteksi kepribadian Big Five pengguna Twitter dan mengukur nilai performansi sistem dengan parameter akurasi.. Data yang digunakan dalam penelitian ini adalah *tweet* yang telah diekstraksi menggunakan metode TF-IDF (*Term Frequency – Inverse Document Frequency*) dan metode LIWC. Hasil ekstraksi fitur menggunakan TF-IDF berupa bobot per-kata, dan dari kumpulan bobot per-kata tersebut dapat diperoleh vektor yang merepresentasikan *tweet* dari pengguna Twitter. Sedangkan, atribut yang diperoleh dari metode LIWC berupa bobot kata-kata yang digunakan pada *tweet* yang mencirikan pengguna Twitter merujuk pada aspek psikologis.

2. Studi Literatur

2.1 Big Five

Dimensi dalam kepribadian Big Five merupakan jenis *trait* kepribadian yang paling baik sebagai acuan dalam sebuah penelitian. *Trait* ini juga baik dalam pengukuran kepribadian seseorang. Pengembangan model ini dengan mengklasifikasikan kata-kata. 5 faktor yang terbentuk dari hasil klasifikasi ini dijadikan penciri dari masing-masing kelompok. Faktor ini kemudian dikenal menjadi 5 *traits* kepribadian Big Five yang meliputi:[8].

a. *Neuroticism*

Neuroticism merupakan jenis *trait* yang melihat dari kestrabilan emosi seseorang. Dari *trait* ini diidentifikasi kecenderungan seseorang untuk mengalami stress. Individu dengan nilai *neuroticism* yang rendah memilki ciri yaitu tenang, bergairah, dan merasa aman. Sedangkan individu dengan nilai *neuroticism* lebih tinggi mudah tertekan, gelisah, dan merasa tidak aman.

b. *Extraversion*

Extraversion merupakan jenis *trait* yang melihat dari intensitas interaksi interpersonal, tingkat ketergantungan terhadap orang lain, dan kemampuan untuk berbahagia. Individu yang memiliki nilai *extraversion* rendah tidak ramah, tenang, menyendiri, *task-oriented*, pemalu, dan pendiam. Individu dengan nilai *extraversion* tinggi lebih mudah bergaul dengan sesama, aktif, banyak berbicara, *person-oriented*, dan selalu optimis.

c. *Openness*

Openness merupakan jenis *trait* yang melihat dari keinginan seseorang dalam mencari dan menghargai pengalaman baru serta lebih senang mengetahui sesuatu yang baru. Individu yang memiliki nilai kepribadian *openness* rendah lebih mengikuti hal yang sudah ada, hanya tertarik pada satu hal saja, jiwa seni yang rendah, dan kurang analitis. Individu dengan nilai kepribadian *openness* tinggi lebih suka akan hal baru, inovatif, kreatif dan imajinatif.

d. *Agreeableness*

Agreeableness merupakan jenis *trait* yang melihat dari kualitas orientasi interpersonal seseorang pada perasaan kasihan dan sikap toleransinya pada suatu hal. Individu dengan nilai kepribadian *agreeableness* rendah lebih sinis, kasar, tidak kooperatif, dan manipulatif. Individu dengan nilai kepribadian *agreeableness* tinggi lebih suka menolong, dapat dipercaya, mudah memaafkan, serta kooperatif.

e. *Conscientiousness*

Conscientiousness merupakan jenis *trait* yang melihat dari kemampuan seseorang dalam berorganisasi, baik mengenai keteraturan maupun motivasi untuk mencapai sebuah. Individu dengan nilai *conscientiousness* rendah lebih malas, sulit untuk dapat dipercaya, tidak disiplin, dan hedonistik. Individu dengan nilai kepribadian *conscientiousness* tinggi lebih teratur, segala seuatu dibuat terencana, dapat dipercaya, pekerja keras, ambisius, dan tekun bekerja[9].

2.2 C4.5

Algoritma C4.5 adalah pengembangan dari algoritma *Iterative Dychotomizer Version 3* (ID3) yang digunakan untuk membuat pohon keputusan dimana pohon keputusan dapat digunakan untuk mengubah sebuah data menjadi pohon keputusan yang menghasilkan aturan-aturan keputusan. Algoritma C4.5 dalam penerapannya digunakan pada pemecahan kasus klasifikasi. Karakteristik dari algoritma C4.5 yaitu pada penentuan nilai *entropy* dan nilai *gain* dari kemungkinan setiap atribut yang menjadi acuan keputusan. Kemudia dilanjutkan dengan proses *ranking* dari keputusan dimana atribut dengan nilai *gain* tertinggi yang akan ditetapkan sebagai *root* dalam *decision tree* sehingga algoritma C4.5 akan digunakan dalam mencari pola keputusan[9].

Pada metode C4.5, sebelum data diekstrak ke dalam bentuk pohon, dilakukan penentuan *root* dari atribut berdasarkan dengan perhitungan entropy kemudian nilai *gain ratio* terbesar, kemudian atribut yang menjadi *internal node* ditentukan untuk setiap *branch* dari *node parent.* Kemudian membuat *node* ketika seleksi atribut dalam kondisi tidak dapat digunakan. Persamaan yang digunakan antara lain[10]:

1. *Entropy*
   Entropy merupakan parameter yang digunakan untuk pengukuran keragaman dari suatu data. Jika nilai *entropy* semakin kecil maka semakin baik digunakan untuk melakukan ekstraksi suatu kelas.

$$Entropy(S) = \sum_{i=1}^{n} - \text{pi} \log_2 \text{pi}$$

   Dimana S merupakan kumpulan data latih, n merupakan jumlah partisi dalam S dan *pi* merupakan proporsi sampel dalam kelas i.
2. *Information gain*
   Information gain menggunakan maksimal *entropy* untuk pembobotan sebuah fitur.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} -\frac{Si}{S} \times Entropy(Si)$$

   Dimana S merupakan kumpulan data latih, A merupakan atribut, n merupakan jumlah partisi dalam atribut A S merupakan jumlah kasus dalam S dan Si merupakan jumlah partisi ke-i.
3. *Gain ratio*
   Gain ratio adalah pengembangan dari *information gain* dimana sebagai perhitungan untuk meminimalisasi bias pada atribut dengan cabang yang banyak.

$$Gain\ ratio = \frac{\text{Gain(S,A)}}{\text{Split Information (S,A)}}$$

   Dimana *split information* sebagai berikut.

$$Split\ Information(S, A) = - \sum_{i=1}^{n} \frac{Si}{S} \log 2 \frac{Si}{S}$$

2.3 Kepribadian Berdasarkan Perilaku Sosial

Perilaku sosial mendefinisikan kepribadian seseorang atau pengguna melalui frekuensi penggunaan media sosialnya[11]. Dalam menunjukkan tingkat perilaku sosial pengguna, terdapat fitur-fitur perilaku sosial berdasarkan penelitian yang dilakukan adalah *follower* yaitu akun twitter lain yang mengikuti akun yang diacu, *following* yaitu pengguna lain yang diikuti oleh pengguna yang diacu, jml_tweet yaitu jumlah pesan yang diunggah oleh pengguna, *retweet* yaitu pengguna mengunggah kembali *tweet* dari pengguna lain, *hastag* menunjukkan bahwa pengguna terlibat pada topik tertentu yang ditandai dengan karakter '#', *mention* yaitu pengguna menyebutkan pengguna lain dalam *tweet*, tanda_baca yaitu simbol dari sebuah kata yang ingin diungkapkan oleh pengguna, huruf_kapital yaitu huruf besar yang digunakan pengguna dalam menulis *tweet*.

2.4 *Term Frequency – Inverse Document Frequency* (TD-IDF)

TF-IDF merupakan metode yang digunakan untuk mengetahui seberapa sering suatu kata muncul dalam dataset dengan memberikan bobot pada setiap kata[12]. Perhitungan TF menggunakan persamaan: $tf = tf_{ij,}$ dengan *tf* merupakan *term frequency*, dan tf_{ij} merupakan jumlah *term* t_i yang muncul dalam dataset $d_{j.}$ perhitungan TF dengan jumlah *term* t_i yang muncul dalam dataset $d_{j.}$ Sedangkan *Inverse Document Frequency* (IDF) dihitung dengan persamaan: $idf_i = log\frac{N}{dfi}$ dengan idf_i merupakan *inverse document frequency*, N merupakan banyaknya dataset diambil oleh system. df_i merupakan jumlah dataset dalam koleksi dimana term t_i tampil di dalamnya. TF-IDF dihitung dengan persamaan: $W_{ij} = tf_{ij} \cdot log\frac{N}{dfi}$. W_{ij} merupakan bobot dataset, N merupakan banyaknya dataset diambil oleh sistem, tf_{ij} merupakan jumlah term t_i yang muncul pada dataset d_j, dan df_i adalah jumlah dataset dalam koleksi dimana *term* t_i tampil di dalamnya[13].

2.5 *Linguistic Inquiry and Word Count* (LIWC)

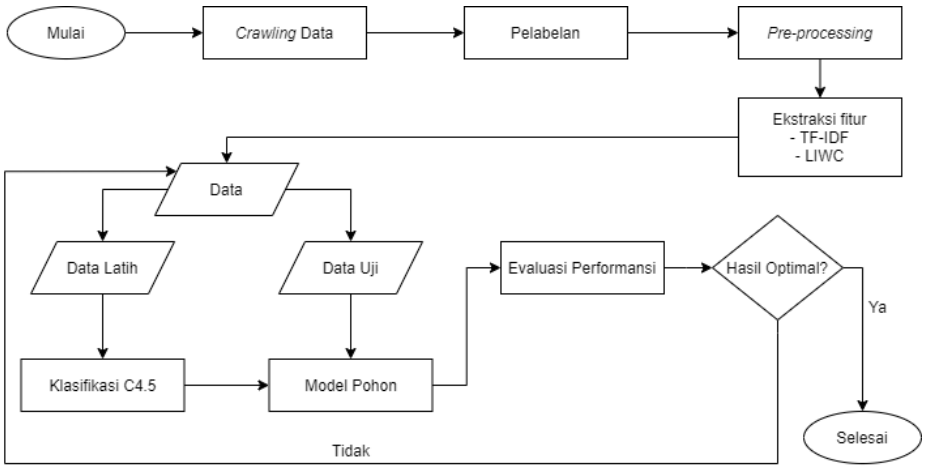
Linguistic Inquiry and Word Count (LIWC) merupakan kakas dimana pengembangannya dilakukan oleh Pennebaker sejak 2007 yang berfungsi sebagai program atau *tools* untuk menghitung kata atau menghitung nilai korelasi secara otomatis berdasarkan *vocabulary* atau kategorinya. LIWC menganalisis penggunaan kata berdasarkan lebih dari 70 kategori, termasuk emosionalitas, kata fungsi (seperti kata benda, kata keterangan dll), masalah pribadi (seperti pekerjaan, uang, dan agama), hubungan sosial dan fokus pada masa[14].

2.6 Metode *Oversampling*

Metode oversampling merupakan metode yang digunakan untuk menyeimbangkan data dengan cara meningkatkan sampel kelas minoritas sampai seimbang atau setara dengan kelas mayoritas dengan menduplikasi sampel kelas minoritas secara acak[15].

3. Sistem Deteksi Kepribadian Dengan C4.5

Sistem yang dibangun untuk prediksi kepribadian *Big Five* pengguna Twitter dengan metode C4.5 dapat dilihat pada Gambar 1.



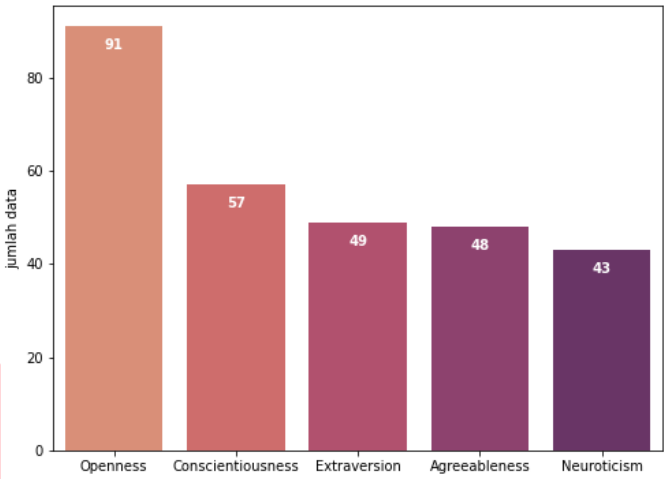
Gambar 1 Sistem Deteksi Kepribadian Dengan C4.5

1. *Crawling* Data
   Dataset yang digunakan pada penelitian ini adalah data hasil *crawling* dari Twitter. Pengumpulan data dengan cara diunduh dari *server* Twitter dengan menggunakan API Twitter[16]. *Crawling* data dilakukan menggunakan sistem crawling data yang telah dibangun oleh peneliti sebelumnya. Atribut-atribut yang digunakan yaitu perilaku sosial dan tweet pengguna. Data pengguna yang akan digunakan untuk menentukan jenis kepribadiannya diambil dari hasil survey menggunakan kuesioner yang telah dibagikan oleh penulis.
2. Pelabelan
   Pelabelan dilakukan dengan cara menyebarkan kuesioner *Big Five Inventory*(BFI) yang telah dikembangkan peneliti sebelumnya dalam bentuk online form yang berisi total 25 pernyataan yang mencirikan setiap jenis kepribadian. Form dibagi ke dalam 5 bagian berdasarkan jenis kepribadian. Pernyataan tersebut berjumlah 5 untuk setiap kepribadian. Setiap pernyataan pada setiap jenis kepribadian dinilai dari skala 1 sampai 5. Nilai yang diperoleh dari perhitungan skala tersebut dibandingkan, jenis kepribadian dengan nilai yang paling tinggi ditentukan sebagai jenis kepribadian dari responden yaitu pengguna twitter.
3. *Pre-processing*
   Pre-Processing data merupakan tahap awal yang digunakan sebelum melakukan langkah-langkah dalam data mining lebih lanjut untuk mendapatkan kualitas data yang lebih baik[17]. Tahapan dalam *pre-processing* data yaitu *Case folding* bertujuan untuk membuat semua text menjadi huruf kecil, *Tokenizing* dalam *pre-processing* merupakan tahap memotong string input sehingga memisah kalimat menjadi kata, *Filtering* merupakan tahap mengambil kata-kata penting dari hasil *tokenizing* dengan *stopword removal*, *Stemming* merupakan tahap mengganti bentuk kata menjadi kata dasar[18].
4. Ekstraksi Fitur
   Teknik yang digunakan dalam ekstraksi fitur yaitu LIWC dan TF-IDF. Pada LIWC, *tweet* setiap pengguna dihitung dengan kamus kategori kata sedangkan pada TF-IDF dilakukan pembobotan pada *tweet*.
5. *Split* Data
   Pada tahap ini, data dibagi menjadi 2 yaitu data latih dan data uji. Rasio skenario data latih dan data uji yaitu (70:30), (80:20) dan (90:10).
6. Metode Klasifikasi C4.5
   Pada tahap ini, dilakukan klasifikasi pada data yang telah dibagi menggunakan metode C4.5.
7. Model Pohon
   Setelah melakukan klasifikasi dengan metode C4.5, dihasilkan model pohon keputusan.
8. Evaluasi Performansi
   Model C4.5 yang telah diperoleh kemudian diuji dengan data uji untuk mencari nilai akurasi sebagai metrik untuk mengevaluasi performansi model C4.5 yang telah dilatih.

4. Evaluasi

4.1 Dataset dan Pelabelan

Jumlah data pada penelitian ini sebanyak 549.151 *tweets* dari 287 akun Twitter dengan total terdapat lima kelas dengan rincian direpresentasikan oleh Gambar 2.



Gambar 2 Proporsi kelas kepribadian pengguna Twitter.

4.2 Perilaku sosial

Perilaku sosial menunjukkan frekuensi penggunaan media sosial twitter oleh pengguna. Fitur pada perilaku sosial yang digunakan dalam penelitian ini adalah jml_follower, jml_following, jml_tweet, website, media_url, retweet, hastag, mention, tanda_baca dan huruf_kapital. Nilai pada setiap atribut dari setiap pengguna ditampilkan pada Tabel 1.

Tabel 1 Dataset perilaku sosial

id	akun	label	jml_follower	jml_following	jml_tweet	website	media_url	Retweet	hastag	mention	tanda_baca	huruf_kapital
0	AttalaRafid_	Extraversion	164	230	4132	52	136	63	9	504	64	2451
1	Clarisa Hasya	Extraversion	80	117	3857	59	450	662	60	936	158	6243
2	Etedadd	Extraversion	148	139	2476	106	188	589	163	1051	293	7976
…	…	…	…	…	…	...	...	…	…	…	…	…
289	zayichi	Openness	730	371	3128	308	94	637	282	2269	1259	16272

4.3 Dataset Tweet

Tweet yang diperoleh dari setiap akun beserta labelnya ditampilkan pada dataset di Tabel 2.

Tabel 2 Dataset tweet

id	akun	label	tweets
0	AttalaRafid_	Extraversion	lucu nih haha nih bocah hari data lessgo bahas...
1	ClarisaHasya	Extraversion	asli benar banget malah jadi buat tidak ada ya...
2	Etedadd	Extraversion	rt are you working with opencv heres cheat she...
…	…	…	…
289	zayichi	Openness	hi how are santai kayak pantai pikir kritis…

4.2 *Pre-processing* dan Ekstraksi Fitur

4.2.1 Hasil *Pre-processing*

Hasil *pre-processing* yang telah dilakukan dimana data yang diperoleh melalui tahap c*ase folding*, *tokenizing*, *filtering* dengan *stopword removal* dan *stemming* dapat dilihat pada sampel di Tabel 3.

Tabel 3 Hasil *Pre-processing*

id	akun	label	tweets	tweets_tokens	tweet_tokens_stemmed
0	AttalaRafid_	Extraversion	lucu nih haha nih bocah hari data lessgo bahas...	['lucu', 'nih', 'haha', 'nih', 'bocah', 'hari'...	['lucu', 'haha', 'bocah', 'data', 'lessgo', 'b...
1	ClarisaHasya	Extraversion	asli benar banget malah jadi buat tidak ada ya...	['asli', 'benar', 'banget', 'malah', 'jadi', '...	['asli', 'banget', 'olah', 'bosan', 'kasihan',...
2	Etedadd	Extraversion	rt are you working with opencv heres cheat she...	['rt', 'are', 'you', 'working', 'with', 'openc...	['are', 'you', 'working', 'with', 'opencv', 'h...
…	…	…	…	…	…
289	zayichi	Openness	hi how are santai kayak pantai pikir kritis…	['hi', 'how', 'are', 'santai', 'kayak', 'di', ...	['hi', 'how', 'are', 'santai', 'kayak', 'panta...

4.2.2 Hasil Ekstraksi Fitur

Tahap ini dilakukan ektraksi fitur dari dataset yang tertera pada Tabel 2 menggunakan dua teknik, yaitu: TF-IDF dan LIWC. Sampel dari hasil ektraksi fitur yaitu bobot pada setiap kata dapat dilihat pada Tabel 4.

Tabel 4 Fitur TF-IDF

id	…	abis	abiss	abistu	able	…
1	…	0.0	0.0	0.0	0.0142	…
2	…	0.0	0.0	0.0319	0.0	…
3	…	0.0	0.0	0.0	0.0	…
…	…	…	…	…	0.0	…
289	…	0.0039	0.0	0.0	0.0168	…

Pada ekstraksi fitur dengan LIWC, menghitung nilai korelasi berdasarkan *vocabulary* atau *tweet* setiap pengguna dihitung dengan kamus kategori dapat dilihat pada Tabel 5.

Tabel 5 Fitur LIWC

id	Sum of orang pertama	Sum of orang kedua	Sum of orang ketiga	Sum of orang jamak	…
1	-8.36	-8.64	-0.6	-0.2	…
2	-17.29	-10.24	-0.78	-1	…
3	-8.55	-5.6	-1.86	-2.5	…
…	…	…	…	…	…
289	-27.74	-35.04	-2.76	-6.3	…

4.3 Hasil Pengujian

4.3.1 Skenario Pertama

Penelitian pada skenario pertama dilakukan terhadap fitur Perilaku Sosial (PS) dengan variasi rasio pembagian data yaitu (70:30), (80:20) dan (90:10) dengan tujuan untuk mendapatkan proporsi data yang terbaik berdasarkan pada metrik akurasi. Hasil skenario pertama dapat dilihat pada Tabel 6.

Tabel 6 Hasil skenario pertama akurasi perilaku sosial (*baseline*)

Rasio Data	Akurasi (%)
	PS(*Baseline*)
70:30	37.93
80:20	**44.82**
90:10	41.37

Dari skenario mencari akurasi dari perilaku sosial, di dapatkan rasio data dengan akurasi tertinggi sebesar 44.82% yaitu 80:20. Rasio data ini akan digunakan untuk skenario selanjutnya.

4.3.2 Skenario Kedua

Penelitian yang dilakukan pada skenario kedua menggunakan rasio data terbaik yang diperoleh dari skenario pertama yang dipadukan dengan salah satu maupun kombinasi dari fitur PS, LIWC dan TF-IDF. Fitur dengan nilai akurasi yang terbaik kemudian diambil untuk digunakan pada skenario berikutnya. Nilai akurasi yang diperoleh dari fitur PS, LIWC, TF-IDF, PS+LIWC, PS+TF-IDF, LIWC+TF-IDF dan PS+LIWC+TF-IDF seperti pada Tabel 7.

Tabel 7 Nilai akurasi fitur skenario kedua

Fitur	Akurasi (%)
PS (Baseline)	44.82
LIWC	44.82
TF-IDF	48.27
PS+LIWC	37.93
PS+TF-IDF	58.62
LIWC+TF-IDF	**62.06**
LIWC+TF-IDF+PS	**62.06**

Dari nilai akurasi yang diperoleh dari pengujian pada skenario kedua seperti ditunjukkan pada Tabel 7, hasil yang diperoleh yaitu akurasi terbesar pada fitur LIWC+TF-IDF dan LIWC+TF-IDF+PS sebesar 62.06%. Hasil pada skenario kedua seperti pada Tabel 8.

Tabel 8 Hasil skenario kedua dengan TF-IDF dan LIWC

Fitur	Akurasi (%)
LIWC+TF-IDF	62.06
LIWC+TF-IDF+PS	62.06

4.3.3 Skenario Ketiga

Pada skenario ketiga dilakukan uji coba penambahan (*over-sampling*) dan pengurangan (*under-sampling*) jumlah data karena proporsi data yang tidak seimbang (*imbalance*). Metode penambahan data yang digunakan adalah *Synthetic Minority Oversampling Technique* (SMOTE), dan metode pengurangan data menggunakan *Random Under-Sampling*. Kemudian, hasil percobaan menggunakan metode *over-sampling* dan *under-sampling* dibandingkan dengan hasil yang diperoleh dari model yang tidak dilakukan manipulasi data. Hasil perbandingan tersebut diperoleh data PS+LIWC+TF-IDF dengan metode *over-sampling* memperoleh akurasi tertinggi yaitu 79.12%. Hasil tersebut dipaparkan pada Tabel 9.

Tabel 9 Hasil skenario ketiga dengan *sampling*

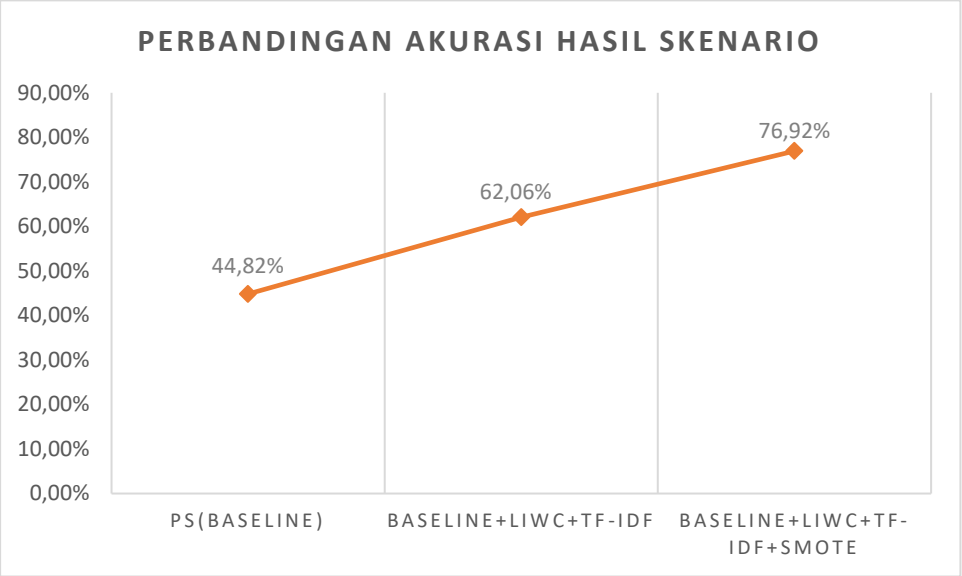
	Akurasi (%)	
	LIWC + TF-IDF	LIWC + TF-IDF + PS
Tanpa *balancing*	62.06	62.06
under-sampling	59.09	40.90
over-sampling	69.23	**76.92**

Pada penelitian ini telah dilakukan tiga macam eksperimen terhadap model pohon keputusan dengan metode C4.5. Pemilihan fitur dan metode yang dilakukan pada skenario penambahan data TF-IDF dan LIWC dan skenario dengan menerapkan metode *sampling* dilakukan berdasarkan pada analisis terhadap hasil yang diperoleh dari skenario pertama yaitu perilaku sosial. Hasil dari ketiga skenario direpresentasikan oleh Tabel 10.

Tabel 10 Perbandingan hasil percobaan

Fitur	Akurasi (%)
PS (Baseline)	44.82
Baseline+LIWC + TF-IDF	62.06 (+17.24)
Baseline+LIWC + TF-IDF + SMOTE	76.92 (+32.1)

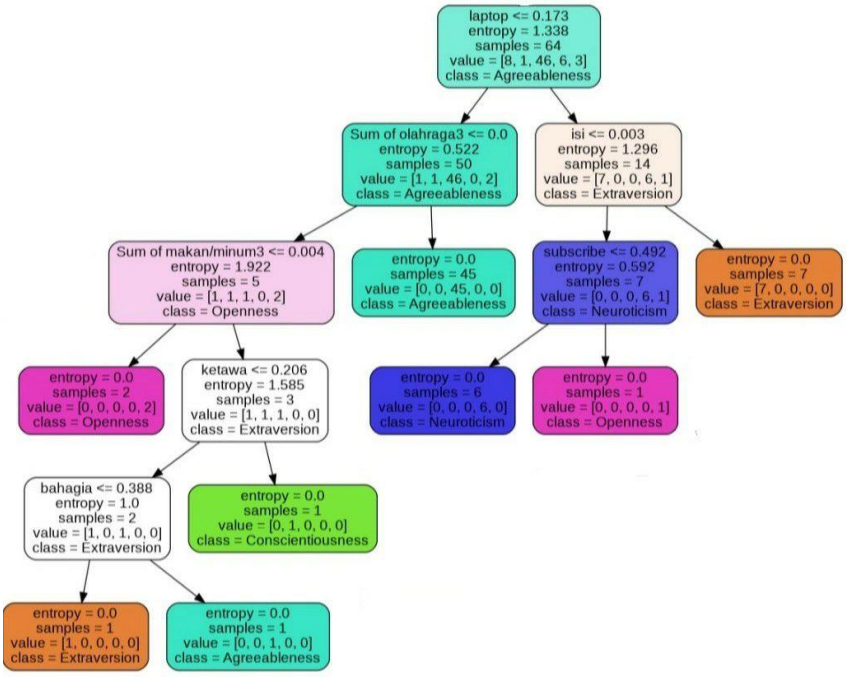
Dari perbandingan hasil percobaan diperoleh data Baseline+LIWC+TF-IDF dengan menerapkan metode SMOTE menghasilkan nilai akurasi tertinggi yaitu 76.92% dimana berarti penambahan akurasi sebesar 32.1%. perbandingan akurasi hasil percobaan atau kenaikan nilai akurasi dari setiap percobaan yang telah dilakukan dengan skenario percobaan ditampilkan pada Gambar 3.



Gambar 3 Grafik Perbandingan Akurasi Hasil Percobaan

4.3.4 Model Pohon

Pohon keputusan yang diperoleh dari data Baseline+LIWC+TF-IDF dengan menerapkan metode SMOTE memiliki jumlah *maximum depth* yang berbeda-beda untuk setiap *decision node*-nya. Pada *root* pohon keputusan terdapat seleksi kondisi terhadap atribut *root* dari masing-masing data, dengan aturan ini maka terbentuklah dua cabang yang untuk memisahkan data yang memenuhi kondisi (*True*) dan data yang tidak memenuhi kondisi (*False*). Pada masing-masing *decision node* terdapat kondisi untuk atribut atau kolom yang lainnya (dari data atribut) dengan nilai aturan yang berbeda pula. Data hasil seleksi pada bagian *root* kemudian diseleksi kembali pada *decision node* di bawahnya, dan berakhir jika *decision node* hanya memiliki satu cabang saja yang dinamakan dengan *termination node*. Melalui *termination node* kepribadian seseorang dapat diprediksi. Contoh pohon keputusan dapat dilihat pada Gambar 4.



Gambar 4 Contoh Pohon Keputusan

Setiap akun pengguna twitter dalam penelitian ini memiliki banyak atribut, untuk memprediksi jenis kepribadian seseorang atau pengguna twitter dilihat dari model pohon yang telah terbentuk yaitu dengan melihat nilai setiap atributnya apakah memenuhi kondisi pada *tree* atau tidak. Atribut awal yaitu laptop memiliki nilai entropy sebesar 1.338 dengan sampel sebanyak 64. Jika akun twitter memiliki atribut laptop yang nilainya kurang dari sama dengan 0.173 maka memenuhi kondisi (*true)*. Jika tidak maka kondisi tidak memenuhi (*false)*. Pengecekan atribut ini dilakukan hingga mencapai *termination node* dimana kondisi berakhir sehingga jenis kepribadian pengguna twitter dapat diketahui. Contohnya jika kita ingin mencari akun dengan jenis kepribadian *Openness*,

dimulai dari pengecekan atribut laptop. Dilakukan pengecekan terhadap isinya, jika terdapat atribut laptop dengan bobotnya memenuhi nilai kurang dari sama dengan 0.173 maka memenuhi kondisi. Setelah diketahui telah memenuhi kondisi, dilanjutkan pada pengecekan atribut Sum of olahraga3 pada decision node ke arah kiri. Dilakukan pengecekan terhadap isinya, jika terdapat atribut Sum of olahraga3 dengan bobotnya memenuhi nilai kurang dari sama dengan 0.0 maka memenuhi kondisi. Setelah diketahui telah memenuhi kondisi, dilanjutkan pada pengecekan atribut Sum of makan/minum3 pada decision node ke arah kiri. Dilakukan pengecekan terhadap isinya, jika terdapat atribut Sum of makan/minum3 dengan bobotnya memenuhi nilai kurang dari sama dengan 0.004 maka memenuhi kondisi dan sampai pada *termination node* dimana jenis kepribadian seseorang diketahui yaitu *Openness.*

5. Kesimpulan

Pada penelitian ini dilakukan pengimplementasian algoritma C4.5 dengan data yang diperoleh dari hasil analisis perilaku pengguna Twitter untuk mendapat model pendeteksi kepribadian Big Five. Tahapan yang dilakukan adalah *crawling* data *tweets* sebanyak 549.151 dari 287 akun Twitter, pelabelan, *pre-processing*, ekstraksi fitur dengan TF-IDF dan LIWC, dan melakukan percobaan dengan tiga skenario yang diujikan menerapkan teknik *hyperparameter tuning* menggunakan Grid Search dengan perilaku sosial sebagai *baseline*. Akurasi yang diperoleh dengan penambahan data TF-IDF dan LIWC mengalami kenaikan sebesar 17.24% dari *baseline* dengan nilai akurasi 62.06% dan menggunakan metode SMOTE dapat meningkatkan nilai akurasi sebesar 32.1% dari *baseline* dengan nilai akurasi 76.92%. Dari fitur PS+LIWC+TF-IDF dengan metode SMOTE, dihasilkan model pohon keputusan deteksi kepribadian Big Five.

REFERENSI

[1] *Psikologi Kepribadian.* (2018). Malang: Universitas Muhammadiyah Malang. Retrieved 10 24, 2020

[2] Adinugroho, S., & Sari, Y. A. (2018). *IMPLEMENTASI DATA MINING MENGGUNAKAN WEKA.* Malang: UB Press.

[3] Claudy, Y. I., Perdana, R. S., & Fauzi, M. A. (2018, Agustus). Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon. *2*. Retrieved Dec 06, 2020, from http://j- ptiik.ub.ac.id

[4] Damanik, A. T., & Khodra, M. L. (2015, Juni). Prediksi Kepribadian Big 5 Pengguna Twitter dengan Support Vector Regression. *3*, 15.

[5] Sidik, M. A. (2019). Sistem Deteksi Kepribadian Berdasarkan Pola Tanda Tangan Menggunakan Metode Support Vector Machine Dan Principal Component Analysis. *Retrieved from http://elibrary.unikom.ac.id/id/eprint/1507*

[6] Utama, B. P. (2018). Support Vector Machine Dalam Sistem Pendeteksi Kepribadian Berdasarkan Pola Tanda Tangan. Retrieved from https://repository.unikom.ac.id/id/eprint/58762

[7] Febrianto, N., Prasetia, I., & Wijaya, A. (2015). Pembuatan Sistem Prediksi Kepribadian "The Big Five Traits" dari. Retrieved 10 24, 2020, from https://ir.cs.ui.ac.id/inacl2016/paper/INACL_2016_paper_3-1-7.pdf

[8] L. A. Pervin and O. P. John. *Personality; Theory and Research.* 8 ed., 2001.

[9] Ganelli, A. E., Dewi, R., Rabialdi, Yusri, & Junaidi. (2010). *Kepribadian perempuan Aceh yang tangguh: kemarin, sekarang, dan esok.* Medan: Usu Press. Retrieved Dec 06, 2020

[10] Herwijayanti, B., Ratnawati, D. E., & Muflikhah, L. (2018, Januari). Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDFdan Cosine Similarity. *2*, 307.

[11] *Psikologi Kepribadian.* (2018). Malang: Universitas Muhammadiyah Malang. Retrieved 10 24, 2020

[12] Jennifer Golbeck, Cristina Robles, Michon Edmondson, Karen Turner. *Predicting Personality with Social Media*. CHI Extended Abstracts. Pp: 253-262, 2011

[13] Sembodo, J. E., Setiawan, E. B., & Baizal, Z. A. (2016, Sept). Data Crawling Otomatis pada Twitter. 12. doi:10.21108

[14] *Yudha Yudhanto, S. (2018). Belajar Mengelola Facebook dan Twitter. Retrieved 10 24, 202*

[15] Heranova, O. (2021). Synthetic Minority Oversampling Techniquepada Averaged OneDependence Estimatorsuntuk Klasifikasi Credit Scoring. Retrieved from http://jurnal.iaii.or.id

[16] Hormansyah, D. S., & Utama, Y. P. (2018, Mei). APLIKASI CHATBOTBERBASIS WEBPADA SISTEM INFORMASI LAYANAN PUBLIK KESEHATAN DI MALANG DENGAN MENGGUNAKAN METODE TF-IDF. *4*, 226.

[17] Permana, S., Rahayu, W. I., & Fatonah, R. N. (2020). *Implementasi Algoritma C4.5 Dalam Penentuan Penerima Bonus Tahunan Pegawai.* (R. M. Awangga, Ed.) *Bandung: Kreatif Industri Nusantara. Retrieved 10 24, 2020*

[18] Permana, S., Rahayu, W. I., Rd.Nuraini, & Fatonah, S. (2020). *Implementasi Algoritma C4.5 Dalam Penentuan Penerima Bonus Tahunan Pegawai.* (R. M. Awangga, Ed.) Bandung: Kreatif Industri Nusantara