Prediksi Kepribadian DISC Pada Twitter Menggunakan Metode Decision Tree C4.5 dengan Pembobotan TF-IDF dan TF-RF

Maulina Gustiani Tambunan[1], Erwin Budi Setiawan[2]

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

Abstrak

Media sosial online berperan penting bagi kebutuhan manusia sehari-hari, yang dapat membangun konektivitas antara pengguna dengan pengguna lainnya. Di zaman yang modern ini hampir setiap kalangan memiliki media sosial sebagai sarana informasi dan sebagai ungkapan dengan sudut pandang keperibadian masing-masing pengguna atas berbagai aspek kehidupan. Dengan menggunakan media sosial di twitter, penulis dapat menemukan berbagai macam karakter dan keperibadian yang dimiliki oleh masing-masing pengguna media sosial. Salah satu permasalahannya adalah bagaimana mengklarifikasi kepribadian seseorang melalui tweet pada media sosial twitter. Maka dalam hal ini, penulis melakukan klasifikasi kepribadian pengguna twitter dengan menggunakan penilaian DISC serta membangun sistem klasifikasi kepribadian untuk mengetahui tingkat akurasi algoritma metode *Decision Tree C4.5* dan nilai performansi yang baik. Yang membedakan penelitian ini dengan penelitian lainnya adalah setiap kata menggunakan metode pembobotan *TF-IDF* dan *TF-RF*. *Followers, following, retweet, hastag*, huruf besar, huruf kecil, *emoticon* dan lainnya merupakan sarana pendekatan berdasarkan prilaku sosial pengguna. Percobaan melalui pendekatan perilaku sosial didapatkan hasil nilai *Bi Class* Dominance adalah sebesar 97.56, Influence sebesar 70.13, Steadiness sebesar 57.14, Compliance sebesar 73.34. Percobaan melalui pendekatan linguistik TF-IDF didapatkan hasil nilai *Bi Class* Dominance adalah 97.50, Influence sebesar 73.17, Steadiness sebesar 46.15, Compliance sebesar 74.36. Percobaan melalui pendekatan linguistik TF-RF didapatkan hasil nilai *Bi Class* Dominance adalah sebesar 100, Influence sebesar 75.00, Steadiness sebesar 53.28, Compliance sebesar 72.50. Hasil perhitungan *F-Measure* dari pendekatan perilaku sosial adalah sebesar 0.2626 dari data latih dan data uji 80 :20, melalui pendekatan linguistik pada TF-IDF didapatkan hasil perhitungan *F-Measure* sebesar 0.3010 dari data latih dan data uji 90 :10, dan percobaan melalui pendekatan linguistik pada TF-RF didapatkan hasil *F-Measure* sebesar 0.4824 dari data latih dan data uji 90 :10.

Kata Kunci : DISC, Decision Tree C4.5, TF-IDF, TF-RF, *Twitter*

Abstract

Online social media plays an important role for everyday human needs, which can build connectivity between users and other users. In this modern era almost every society has social media as a means of information and as an expression with each user's personal perspective on various aspects of life. By using social media on Twitter, the author can find a variety of characters and personality possessed by each social media user. One problem is how to clarify one's personality through tweets on Twitter social media. So in this case, the writer classifies the personality of Twitter users by using the DISC assessment and builds a personality classification system to determine the accuracy of the Decision Tree C4.5 algorithm algorithm and good performance value. What distinguishes this study from other studies is that each word uses the TF-IDF and TF-RF weighting methods. Followers, following, retweet, hashtag, uppercase, lowercase, emoticons and others are a means of approach based on the user's social behavior. Experiments through the approach of social behavior obtained the value of Bi Class Dominance is 97.56, Influence is 70.13, Steadiness is 57.14, Compliance is 73.34. Experiments through the TF-IDF linguistic approach showed that the Bi Class Dominance value was 97.50, Influence was 73.17, Steadiness was 46.15, Compliance was 74.36. Experiments through the TF-RF linguistic approach showed that the Bi Class Dominance value was 100, Influence was 75.00, Steadiness was 53.28, Compliance was 72.50. The F-Measure calculation results from the social behavior approach is 0.2626 from the training data and 80: 20 test data, through the linguistic approach on TF-IDF, the F-Measure calculation results are obtained from 0.3010 from the training data and 90: 10 test data, and experiments through linguistic approach to TF-RF obtained F-Measure results of 0.4824 from training data and 90: 10 test data.

Keywords : DISC, Decision Tree C.45, TF-IDF, TF-RF, Twitter

1. Pendahuluan

Media sosial adalah media berbasis internet yang memungkinkan pengguna berkesempatan untuk berinteraksi dan mempresentasikan diri, baik secara seketika maupun tertunda, dengan khalayak luas maupun tidak yang mendorong nilai dari *use-generate content* dan presepsi interaksi dengan orang lain. (Caleb T. Carr dan Rebecca A. Hayes, 2015).

Salah satu media sosial yang paling banyak digunakan saat ini adalah media sosial twitter. Oleh karena itu penulis menggunakan twitter sebagai bahan penelitian untuk memprediksi kepribadian para pengguna twitter berdasarkan tweet pengguna. Salah satu alasannya adalah banyak orang yang menggunakan twitter sebagai sarana untuk berinteraksi, mencurahkan perasaan atau suasana hatinya, menyampaikan pendapat atas suatu permasalahan yang terjadi. Dengan informasi yang terdapat didalam fitur media sosial twitter dapat mencerminkan kepribadian penggunanya.

Untuk melakukan tes kepribadian dapat dilakukan dengan menggunakan metode DISC, MBTI, Strength Finder dan Big Five. Dalam penelitian Tugas akhir ini, peneliti menggunakan metode prediksi kepribadian DISC, karena model kepribadian DISC berkonsentrasi pada preferensi perilaku sehingga lebih diterapkan, jelas dan mudah dipahami disbanding dengan MBTI [24]. Kepribadian DISC adalah test penilaian kepribadian yang sederhana untuk memahami tipe-tipe perilaku dan gaya kepribadian seseorang. DISC terdiri dari 4 tipe kepribadian yaitu : Dominance (D), Influence (I), Steadiness (S), Compliance (C).

Dalam menganalisi kepribadian DISC melalui pengguna twitter dibutuhkan metodologi yang tepat untuk mendapatkan hasil yang akurat dengan menggunakan metode klasifikasi Decision Tree C4.5 dengan dua pembobotan yaitu TF-IDF dan TF-RF. Pada penelitian sebelumnya [2] telah melakukan penelitian dengan menggunakan klasifikasi Decision Tree C4.5 yang mengatakan bahwa algoritma C4.5 dainggap sebagai algoritma yang sangat membantu dalam melakukan klasifikasi data karena karakteristik data yang diklasifikasi dapat diperoleh dengan jelas, baik dalam bentuk struktur pohon keputusan (Decision Tree) maupun dalam bentuk aturan atau rule if-then sehingga dapat memudahkan pengguna dalam melakukan penggalian informasi terhadap data yang bersangkutan.

Maka pada Tugas akhir ini penulis melakukan penilian untuk mendapatkan tingkat akurasi yang tinggi dari algoritma lainnya, melalui pendekatan perilaku sosial dan pendekatan linguistik menggunakan pembobotan nilai TF-IDF dan TF-RF serta membandingkan nilai akurasi pada kedua pembobotan tersebut dan mendapatkan fitur-fitur dan bentuk model yang terbaik. Yang membedakan penelitian ini dengan penelitian lainnya adalah menggunakan metode pembobotan *TF-IDF* dan *TF-RF* untuk setiap kata tweet pengguna seperti *Followers, following, retweet, hastag*, huruf besar, huruf kecil, *emoticon* dan lainnya merupakan sarana pendekatan berdasarkan prilaku sosial pengguna.

2. Studi Terkait

2.1 Kepribadian DISC

Kepribadian merupakan salah satu metode yang dikenal dalam dunia psikologi untuk menginterpretasi kepribadian seseorang, terutama untuk menemukan hubungan kepribadian dengan lingkungan. Kepribadian seseorang dapat dilihat dari berbagai aspek yaitu kepribadian *openness, extraversion* dan lain-lainnya. Pengukuran penilaian kepribadiaan dapat dilakukan dengan menggunakan beberapa metode seperti MBTI, DISC, *strength finder* dan *Big Five*.

DISC dikemukakan oleh ahli psikolog asal amerika yang bernama William Moulton Marston pada tahun 1928 dalam bukunya yang berjudul *Emotions Of Normal People.* Ia berteori bahwa ekspresi perilaku emosi bisa dikategorikan menjadi 4 tipe perilaku individu ketika berinteraksi dengan lingkungannya yaitu : *Dominance (D), Influence (I), Steadiness (S),* dan *Compliance (C).* [15]

a. *Dominance*
   Karakteristik orang tipe D antara lain tegas, ambisius, independen, menyukai persaingan, penerima tantangan, cepat dalam mengambil keputusan, penuntut, tidak sabar, dan tidak menyukai hal yang rutin.
b. *Influence*
   Karakteristik orang tipe I antara lain ramah, senang bergaul, suka menghibur orang lain, antusias, optimis, motivator, kurang memerhatikan detail, banyak bicara, mudah lupa, dan seringkali bereaksi berlebihan terhadap sesuatu.
c. *Steadiness*
   Karakteristik orang tipe S antara lain sabar, gigih, jujur, akomodatif, loyal, tidak terlalu menuntut, ingin menolong orang lain, tidak suka dengan perubahan, kurang antusias, kurang tegas, cenderung menghindar dari konlik, dan sulit menyusun prioritas.
d. *Compliance*
   Karakteristik orang tipe C antara lain teliti, terstruktur, berhati-hati dalam membuat keputusan, kritis dalam menganalisa kerja sendiri maupun kerja kelompok, patuh terhadap atasan/pimpinan, kurang fleksibel, defensif ketika dikritik, terlalu mengikuti aturan, dan lamban dalam menyelesaikan tugas karena terlalu memperhatikan detail dan menginginkan kesempurnaan.

2.2 Penggunaan Fitur

Pada penelitian ini penulis melakukan analisa melalui pendekatan terhadap perilaku sosial pengguna *twitter* dan pendekatan linguistik atau penggunaan bahasa atau kata yang digunakan pengguna *twitter* pada saat menuliskan *tweet.*

2.2.1 Kepribadian Berdasarkan Pendekatan Perilaku Sosial

Perilaku sosial mendefinisikan kepribadian melalui frekuensi penggunaan media sosial dan tingkat keaktifan antar pengguna. Fitur yang menunjukkan tingkat perilaku sosial pengguna *Twitter* berdasarkan penelitian yang dilakukan [17] adalah sebagai berikut.

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

Selain fitur di atas, terdapat fitur dari *twitter* yang dapat dijadikan bahan pertimbangan untuk dilakukan analisis terkait fitur yang menunjukkan tingkat keaktifan perilaku sosial pengguna *twitter* yaitu sebagai berikut.

a. Jumlah retweet adalah banyaknya pengguna mengunggah kembali *tweet* dari pengguna lain.
b. Jumlah media URL adalah banyaknya tautan berupa gambar atau video yang diunggah oleh pengguna.
c. Jumlah like adalah banyaknya like (digambarkan dengan hati kecil) yang digunakan untuk memberikan apresiasi pada suatu tweet.
d. Jumlah tanda baca adalah banyaknya simbol sebuah dari sebuah kata yang ingin diungkapkan oleh pengguna, tanda baca yang dihitung adalah tanda tanya (?) dan tanda seru (!).
e. Jumlah emoji adalah banyaknya karakter unik yang dapat digunakan oleh pengguna saat menulis *tweet*-nya untuk menggambarkan emosi pengguna melalui karakter-karakter unik. Emoji yang diambil dari link [17] disimpan untuk dimasukan kedalam kamus pada *database.* Total emoji yang didapat berjumlah 2.552 karakter.
f. Jumlah huruf besar adalah banyaknya huruf kapital yang digunakan pengguna saat menulis *tweet.*
g. Kepemilikan bio adalah ada atau tidaknya biografi atau ulasan singkat tentang profil seorang pengguna *twitter* pada laman akun.

2.2.2 Kepribadian Berdasarkan Pendekatan Linguistik

Pada pendekatan Linguistik, dilakukan pencarian atribut berupa penggunaan kata pada *tweet* yang telah dikumpulkan dan digunakan untuk menemukan hubungan antar kata dengan kepribadian seorang 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 mengurai *tweet* ke dalam satuan kata dengan pendekatan unigram. Setelah diurai, satuan kata tersebut diberi bobot dengan perhitungan TF-IDF, dan TF-RF.

2.3 Pre-Processing

Pre-Processing data merupakan langkah-langkah dalam mengolah data mentah yang berikutnya akan dimasukan ke dalam sistem klasifikasi. Proses ini bertujuan untuk menyiapkan data yang akan digunakan secara efisien ke dalam sistem klasifikasi [13] Lalu tahap post-processing yang mencakup semua operasi yang membuat hasil data mining mudah diimplementasikan dan diakses oleh analisis [20]. Beberapa langkah preprocessing dilakukan untuk membersihkan data yang berisi dan akhirnya kepribadian pengguna dipetakan pada DISC untuk mengetahui karakteristik kepribadian.

a. *Case Folding*, merupakan tahapan yang mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf "a" sampai dengan huruf "z" yang diterima. Karakter selain huruf dihilangkan dan dianggap delimeter (pembatas) (Triawati, 2009).

b. *Tokenization*, adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Selain itu, spasi digunakan untuk memisahkan antar kata.
c. *Filtering*, adalah tahap mengambil kata-kata penting dari hasil tokenizing. Proses filtering dapat menggunakan algoritma stoplist (membuang kata yang kurang penting) atau wordlist (menyimpan kata penting). Stoplist atau stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words.
d. *Stemming*, suatu proses yang terdapat dalam sistem IR yang mentransformasi kata-kata yang terdapat dalam suatu dokumen ke kata-kata akrnya (root word) dengan menggunakan aturan-aturan tertentu. Stremming kebanyakan digunakan pada teks berbahasa inggris karena teks berbahasa inggris memiliki struktur imbuhan yang tetap dan mudah untuk diolah. Stemming untuk proses berbahasa Indonesia memiliki struktur imbuan yang rumit atau kompleks sehingga lebih sulit utuk diolah.

Tabel 1 Contoh Alur *Prepocessing*

Awal	Sesudah Case folding	Sesudah tokenizing	Sesudah Filtering	Sesudah Stemming
Bagaimana cara menenangkan HATI	bagaimana cara menenangkan hati	“bagaimana”, “cara”, “menenangkan”, “hati”	“cara”, “menenangkan”, “hati”	“cara”, “tenang”, “hati”

2.4 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. 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..

2.4.1 *Term Frequency Inverse Document Frequency*

Metode TF-IDF merupakan metode untuk menghitung bobot setiap kata yang paling umum digunakan pada information retrieval. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat [9]. Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Metode TF-IDF menggunakan dua konsep yaitu frekuensi kemunculan sebuah kata di dalam sebuah dokumen dan inverse frekuensi dokumen yang mengandung kata tersebut [25].TF-IDF dapat ditulis dengan persamaan:

$$tfidf_t = f_{t,d} \times \log\frac{N}{df_t} \quad (1)$$

Dimana :

$tfidf_t$: bobot dari term *t*

$f_{t,d}$: frekuensi munculnya *term t* pada dokumen *d*

N : jumlah kumpulan dokumen

df_t : jumlah dokumen yang mengandung *term t*

2.4.2 *Term Frequency – Relevan Frequency*

Relevan frequency adalah pemberian bobot dengan mempertimbangkan relevansi dari dokumen untuk term yang sering muncul dalam kaitan tertentu. Sehingga RF hanya akan meningkatkan weight dari terms kategori positif sementara bobot terms dengan kategori negatif akan semakin berkurang, dengan kata lain RF mendiskriminasi terms yang sering muncul dalam kategori negative [27]. Persamaannya sebagai berikut:

$$TF \times RF(t,c) = TF\,(d,t) \times log(2\,\frac{A}{max(1,C)}) \quad (2)$$

Dimana:

A : Jumlah dokumen *tweet* dalam kelas c yang mengandung term t

C : Jumlah dokumen *tweet* bukan kelas c yang mengandung term t

c : Kelas Kategori
t : Kata

2.5 Klasifikasi

Kalsifikasi adalah proses dari mencari suatu himpunan model (fungsi) yang dapat mendeskripsikan dan membedakan kelas-kelas data atau konsep-konsep, dengan tujuan dapat menggunakan model tersebut untuk memprediksi kelas dari suatu objek yang mana kelasnya belum diketahui. Terdapat dua jenis dalam klasifikasi teks, yaitu klasifikasi teks *supervised* dan klasifikasi teks *unsupervised* [24]. Klasifikasi teks *supervised* merupakan proses klasifikasi teks dengan metode learning pada dokumen yang berlabel kelas pada data latih untuk dokumen pembelajaran. Sedangkan untuk klasifikasi teks *unsupervised* merupakan metode klasifikasi teks yang diterapkan tanpa menggunakan label kelas pada data latih untuk menganalisis hubungan antar data [24]. Dalam penelitian tugas akhir ini, jenis klasifikasi teks yang dilakukan penulis termasuk ke jenis *supervised*, karena proses pembelajaran pada dokumen data latih yang memiliki label kelas. Salah satu metode klasifikasi *supervised* adalah *Decision tree C4.5*

2.6 Decision Tree C4.5

Metode terkait klasifikasi supervised yaitu decision tree. Decision tree adalah teknik data mining untuk menyelesaikan masalah klasifikasi dengan cara mengamati beberapa sampel untuk mengetahui bagaimana perilaku vektor atribut tersebut. Dasar dari algoritma ini adalah Greedy algorithm yang membangun pohon keputusan dengan cara rekursif devide and conquer top-down. Jenis decision tree yang digunakan adalah decision tree C4.5 karena karakterisitik data yang digunakan bersifat kontinu. Beberapa pengembangan yang dilakukan pada C4.5 adalah sebagai antara lain bisa mengatasi missing value, bisa mengatasi continue data, dan pruning [21]. Artikel jurnal di [8] menyebutkan algoritma C4.5 merupakan salah satu teknik decision tree yang sering digunakan, yang menghasilkan beberapa aturan-aturan dan sebuah pohon keputusan dengan tujuan untuk meningkatkan keakuratan dan prediksi yang sedang dilakukan, disamping itu algoritma C4.5 merupakan algoritma yang mudah dimengerti.

1. Entropy

Entropy adalah suatu parameter untuk mengukur heterogenitas dari suatu data. Semakin kecil nilai Entropy maka semakin baik untuk digunakan dalam mengekstraksi suatu kelas.

$$Entropy(S) = \sum_{i=1}^{n} - p_i \, log_2 \, p_i \tag{3}$$

Dimana:

S : Kumpulan data latih n: Jumlah partisi dalam S p : proporsi sampel dalam kelas i

2. Information Gain

Nilai yang didapat dari perhitungan *entropy* masih belum asli tetapi, pengukuran efektivitas atribut dalam mengklasifikasikan data latih dapat ditentukan dengan informasi yang telah diperoleh. Dengan persamaan.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{Si}{S} \times Entropy(Si) \tag{4}$$

Dimana:

S : Kumpulan data latih A : atribut n : Jumlah partisi dalam atribut A Si : Jumlah partisi ke-i

3. Gain Ratio

Gain Ratio merupakan modifikasi dari information gain untuk mengurangi bias atribut yang memiliki banyak cabang. Dimana persamaan ditulis sebagai.

$$Gain\ ratio = \frac{Gain(S, A)}{Split\ Information\ (S, A)} \tag{5}$$

Dimana *split Information* sebagai berikut.

(6)

$$Split\ Information(S,A) = -\sum_{i=1}^{n} \frac{si}{s} log_2 \frac{si}{s}$$

2.7 Evaluasi Performansi

Dalam tahapan pengukuran performansi adalah tahap analisis dan evalusai pada sistem yang akan dirancang. Dalam penelitian tugas akhir ini, digunakan performansi yang diukur dengan menggunakan nilai akurasi, *precision* , dan *recall*. Untuk mempermudah menghitung performansi maka digunakan *confusion matrix*.

Table 2 *confusion matrix*

	Predicted Class		
Actual Class		*class = yes*	*class = no*
	class = yes	TP	FN
	class = no	FP	TN

Dimana :

a. TP (*True Positive*) : Kelas yang diprediksi yes, dan ternyata faktanya yes (hasil yang benar).
b. TN (*True Negative*) : Kelas yang diprediksi no, dan ternyata faktanya no (tidak adanya hasil yang benar).
c. FP (*False Positive*) : Kelas yang diprediksi yes, tetapi faktanya no (hasil yang tidak diharapkan).
d. FN (*False Negative*) : Kelas yang diprediksi no, tetapi faktanya yes (hasil yang meleset).

a. 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 [14].

$$Akurasi = \frac{TP + TN}{(TP + FP + TN + FN)} \quad (7)$$

b. Precision

Precision merupakan rasio dari jumlah ketepatan prediksi suatu kelas terhadap jumlah total prediksi yang diklasifikasikan ke dalam kelas tersebut. Berikut rumus dari *precision*:

$$Precision\ (P) = \frac{TP}{(TP+FP)} \quad (8)$$

c. Recall

Recall merupakan rasio dari jumlah ketepatan prediksi suatu kelas terhadap jumlah total fakta yang diklasifikasikan ke dalam kelas tersebut. Berikut rumus dari *recall*:

$$Recall\ (R) = \frac{TP}{(TP+FN)} \quad (9)$$

Untuk menggabungkan rumus *precision* dan *recall* menjadi sebuah rumus tunggal yang disebut *F-Measure* atau *F-1 Score*, dapat dihitung dengan menggunakan rumus berikut :

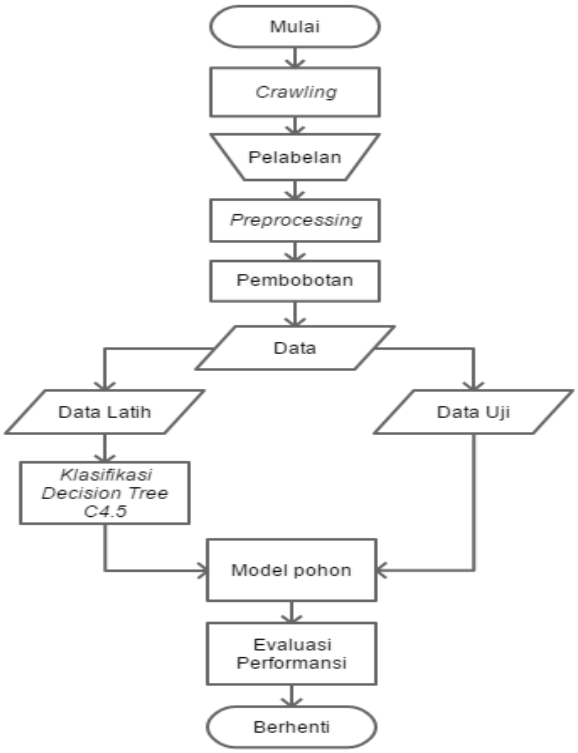
d. F-1 Score

$$F-1\ SCORE = 2\ \frac{\text{Precision x Recall}}{\text{Precision+Recall}} \quad (10)$$

3. Sistem yang Dibangun

Dilakukan perancangan sistem yang akan dibangun untuk mengklasifikasi kelas masing-masing menggnakan metode Decision Tree *C4.5* untuk memprediksi kepribadian DISC. Data yang digunakan adalah dari postingan *tweets* pengguna twitter yang didapatkan melalui *crawling data* menggunakan API twitter. Hasil yang diharapkan adalah model pembelajaran yang dapat mengklasifikasikan kepribadian pengguna twitter.

Rangkaian proses dalam sistem ini akan dibangun seperti Gambar 1.



Sistem Gambar 1 Rancangan

1. *Crawling Data*
   Data yang digunakan pada penelitian ini berupa hasill *tweet* dan beberapa atribut kebutuhan dari fitur yang akan dianalisis. Setiap *crawling data tweet* yang diambil di setiap akunnya bisa sampai 3200 *tweet* terbarunya. Akun yang dapat di*crawling* sejumlah 415 akun. Akun tersebut dicari berdasarkan pencarian akun *twitter* mahasiswa yang mengikuti tes psikotes Telkom University 2013. Mesin *crawling* dapat mengumpulkan data sebanyak 863826 *tweet* dari total keseluruhan akun. Data yang diambil akan dijadikan acuan oleh sistem dan dijadikan *data training* serta *data testing.*

2. Pembagian Data
   Pada tahap ini data yang sudah dikumpulkan akan dipisahkan menjadi data latih dan data uji. Jumlah rasio pembagian data dicoba dengan beberapa skenario demi mengetahui pada rasio data latih dan data uji berapa model dapat berperformansi dengan baik. Pada penelitian ini penulis membuat skenario pembagian data, yaitu: data latih dan data uji (70:30), data latih dan data uji (80:20), dan data latih dan data uji (90:10).

3. *Pre-processing*

Pada tahap ini, *data training* dan *data testing* akan dilakukan *preprocessing data* untuk menghilangkan data yang tidak sempurna. Beberapa tahapan diantaranya ialah *case folding, tokenizing, filtering* dan *stemming*.

a. *Case folding*, yaitu mengubah seluruh huruf kapital menjadi huruf kecil.
b. *Tokenizing*, yaitu mengubah kalimat menjadi kumpulan satu kata.
c. *Filtering*, yaitu menghilangkan *stop word*.
d. *Stemming*, yaitu mengembalikan kata ke dalam bentuk dasar (kata dasar) dengan menghilangkan aditif yang ada.

4. Pembobotan TF-IDF dan TF-RF
Pada proses ini data *tweet* yang telah dikumpulkan sebelumnya akan di proses untuk dilakukan pembobotan yang bertujuan untuk mendapatkan rating pada kata–kata yang didapatkan. Dua metode pembobotan ini akan dicoba untuk mengetahui seberapa berpengaruh kata dari suatu dokumen nantinya. *Term* yang dipilih untuk dijadikan atribut ditentukan ke dalam beberapa skenario pada 45 kata, 30 kata dan 15 kata yang tersering muncul pada setiap kelas nya. Hasil dari pembobotan berupa banyak kata yang telah diberi bobot.

5. Klasifikasi Decision Tree C4.5
Pada proses ini data yang sudah di *preprocessing* akan masuk pada tahap klasfikasi. Dimana data latih diinputkan dan dihitung berdasarkan proses *Decision Tree* C4.5 dengan perhitungan nilai *entropy, information gain, split info* dan nilai *gain ratio.* Maka selanjutnya output dari proses data latih yang sudah di inputkan adalah model prediksi dari data test yang akan membantu evaluasi performansi dari model prediksi yang sudah didapatkan

6. Evaluasi Performansi
Pada tahapan evaluasi performansi ini akan dilakukan perhitungan akurasi, *precision,* dan *recall* dimana untuk mengukur performansi dari sistem yang sudah dibuat.

4. Hasil Analisa dan Uji

Pada bagian ini akan dijelaskan bagaimana hasil uji dari sistem yang telah dibangun sesuai dengan flowchart yang telah dibuat sebelumnya, serta akurasi, *precision*, *recall* yang didapat.

4.1. Data Set

Pada data hasil dari test psikotes mahasiswa Universitas Telkom 2013, didapatkan total akun pengguna yang dapat di *crawling* hanya berjumlah 415 akun dari total peserta lebih kurang 1600 peserta. Tidak tecapainya jumlah akun sesuai dengan totap peserta dikarenakan adanya peserta yang tidak memiliki akun *twitter*, akun dalam keadaan privasi dan ada yang sudah ganti nama akun.

Setiap peserta yang mengikuti hasil tes psikotes akan mendapatkan hasil berupa kepribadian yang paling dominan hingga yang paling tidak dominan. Beriukut merupakan beberapa contoh dari data tes pesikotes mahasiswa Universitas Telkom 2013.

Tabel 3 Data Hasil Psikotes Mahasiswa Universitas Telkom 2013

No	Nama Mahasiswa	DISC	Keterangan
1	ANNISA UTARI	IS/DC	
2	BELLA CITRA HADINI	0	TIDAK HADIR
3	CATUR DHARMA RAMADHAN TRI PUTERA	CSI/D	
4	DENISHA OCTAVIA	SIC/D	
5	DIAN SULISTIYARINI	CSI/D	
…	…	…	…

Dari tabel diatas didapatkan setiap kepribadian mahasiswa dari yang paling dominan melalui huruf pada label yang paling pertama. Dari data tersebut akan diambil sifat yang paling dominan sebagai kelas, seperti yang ada pada mahasiswa pertama yaitu memiliki label kepribadian IS/DC maka kepribadian yang diambil sebagai

kelas ialah I. Diambilnya kepribadian yang paling dominan dikarenakan peneliti hanya memakai 4 kategori untuk melakukan prediksi kepribadian DISC, kelas 1 merupakan kepribadian D (*Dominance*), kelas 2 merupakan kepribadian I (*Influence*), kelas 3 merupakan kepribadian S (*Steadiness*), dan kelas 4 merupakan kepribadian C (*Compliance*).

Tabel 4 Data Hasil Crawling Akun Twitter Peserta Psikotes

ID	Nama	Akun	Follower	Following	Retweet	Hashtag	…	Label
3	Fauzan Abdurrahman	siojanpaujan	34	187	92	151	…	**Compliance**
4	Gandung Tarispranoto	Gandungtaris	4	0	0	0	…	**Steadiness**
5	Hilmy M.F	hilmyncek	313	262	231	275	…	**Influence**
11	Sheyla Putri Nevertari	Sheylaputnev	844	481	532	161	…	**Dominance**

4.2. Hasil Uji

Pada bagian ini akan dijelaskan hasil uji dari sistem yang telah dibuat dan dirancang sesuai dengan flowchart yang telah dibuat sebelumnya, serta akurasi dan kompetensi pengguna twitter yang didapat.

4.2.1 Hasil Preprocessing dan Pembobotan

Pada tahap ini, semua tweet akan dikumpulkan menjadi satu bagian untuk setiap akun dan setelah itu akan dilakukan *preprocessing* untuk mengubah data menjadi terstruktur. Proses *preprocessing* yang pertama adalah penghapusan *retweet* dari tweet, kemudian diikuti dengan penghapusan URL. *Case folding* yaitu mengubah huruf besar menjadi huruf kecil. *Tokenizing* merupakan proses penghapusan karakter seperti titik (.) dan koma (,) lalu, *tweet* diuraikan menjadi satuan kata. *Filtering* yaitu pemilihan kata-kata penting setelah proses *tokenizing*. Terakhir adalah *stemming* yaitu proses mengubah kata yang berimbuhan menjadi kata dasar.

Tabel 5 Contoh Hasil Preprocessing

NAMA: Aisyah asti averossy	**Sebelum Preprocessing**	“Hahahah iya jeng, tpi rada bingung pakenya" "@Tangerang"
	Setelah Preprocessing	Iya bingung

Selanjutnya masuk ke tahap pembobotan, semua kata hasil *preprocessing* akan di beri bobot dengan menggunakan perhitungan TF-IDF dan TF-RF. Berikut adalah hasil dari pembobotan dengan menggunakan perhitungan TF-IDF dan TF-RF.

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

Akun	Kata						
	lama	**Rasa**	**Harap**	**Bias**	**ajar**	**Ingin**	**Sikap**
TF 1	20	24	4	32	6	3	1
TF 2	**61**	**31**	**6**	**118**	**15**	**7**	**0**
TF 3	**40**	**46**	**10**	**146**	**40**	**9**	**0**
…							
IDF 1	**1.0952**	**0.9928**	**2.028**	**0.5372**	**1.3236**	**1.6744**	**6.2972**
RF 1	**4.9694**	**4.9737**	**4.9768**	**4.9647**	**4.9745**	**4.9752**	**4.96663**
…							
TF_IDF 1	**21.904**	**23.8272**	**8.112**	**17.1904**	**5.2944**	**5.0232**	**6.2972**
TF_IDF 2	**66.8072**	**30.7768**	**12.168**	**63.3896**	**19.854**	**11.7208**	**0**
TF_IDF 3	**43.808**	**45.6688**	**20.28**	**78.4312**	**52.944**	**15.0696**	**0**
…							
TF_RF 1	**99.388**	**119.3688**	**19.9072**	**158.8704**	**19.898**	**14.9256**	**4.9666**
TF_RF 2	**303.1334**	**154.1847**	**29.8608**	**585.8346**	**74.6175**	**34.8264**	**0**
TF_RF 3	**198.776**	**228.7902**	**49.768**	**724.8462**	**198.98**	**44.7768**	**0**

4.2.2 Decision Tree C4.5

Setelah melalui tahap *preprocessing* dan pembobotan akan masuk ke tahap klasifikasi *Decision tree C4.5*. Pada decision tree C4.5 untuk menentukan akar dilihat dari gain ratio tertinggi hingga semua atribut dihitung yang akan menghasilkan aturan pohon. Langkah pertama adalah mengubah atribut kontinu menjadi diskret atau biasa disibut discretization. Metode discretization yang digunakan adalah Segmentation by Natural Partitioning yaitu membagi atribut kedalam beberapa segment interval. Selanjutnya, hitung nilai entropy Total. Untuk menentukan atribut yang akan dijadikan akar maka, hitung nilai entropy setiap atribut, misal atribut “makan > 6” maka hitunglah entropy atribut “makan > 6” dan hitung entropy “makan <= 6” setelah itu hitung information

gainnya. Selanjutnya hitung nilai gain ratio tetapi sebelum menghitung gain ratio harus menghitung split info terlebih dahulu. Jika semua sudah terhitung lihatlah nilai tertinggi pada gain ratio, nilai tertinggi inilah yang menjadi akar pohon untuk menentukan aturan pohon. Ulangi langkah-langkah ini hingga semua atribut dihitung dan akar (root) terbentuk

Total data = 415
Total Dominance = 16
Total Influence = 108
Total Steadiness = 171
Total Compliance = 120
Entropy All = 1.731

Attribute Value	Total Data	Total Dominance	Total Influence	Total Steadiness	Total Compliance	Entropy	Gain
lama <= 79.5104	365	16	96	149	104	1.748	
lama > 79.5104	50	0	12	22	16	0	0.194
rasa <= 94.5003	409	15	107	169	118	1.725	
rasa > 94.5003	6	1	1	2	2	1.918	0.003
bisa <= 268.0938	140	5	36	58	41	1.721	
bisa > 268.0938	275	11	72	113	79	1.736	0
ajar <= 49.745	309	13	83	122	91	1.75	
ajar > 49.745	106	3	25	49	29	1.663	0.003
hidup <= 34.7732	224	10	59	95	60	1.741	
hidup > 34.7732	191	6	49	76	60	1.714	0.002
baik <= 89.505	377	13	99	157	108	1.717	
baik > 89.505	38	3	9	14	12	1.837	0.003
jangan <= 99.432	415	16	108	171	120	1.731	
jangan > 99.432	0	0	0	0	0	0	0
hati <= 79.4928	364	14	96	146	108	1.737	
hati > 79.4928	51	2	12	25	12	1.67	0.002
tak <= 49.738	309	12	81	119	97	1.743	
tak > 49.738	106	4	27	52	23	1.663	0.008
ya <= 551.0817	300	8	71	129	92	1.678	

Gambar 2 Perhitungan Decision Tree C4.5

Dari hasil perhitungan didapat aturan pohon sebagai berikut.

Tabel 7 Salah Satu Contoh Aturan Pohon yang Didapat

No.	Rules	Decision
1	ya <= 575.9052 & sama > 337.6948 & hidup > 34.7732 & bukan > 109.241 & iya > 193.9353 & bias <= 278.0232 & bukan <= 109.241 & tau <= 178.7868 & lupa <= 54.7239 & hati > 79.4928 & harus > 84.4883 & sayang <= 34.8054	*Steadiness*
2	Sayang > 34.8054 & masih > 119.1768 & enak >54.6821 & bias <= 278.0232	*Influence*
3	Mau > 327.7824 & enak > 54.6821 & bias > 278.0232 & sama <=337.6948	*Compliance*
4	Cinta <= 44.7543 & cinta > 44.7543 & enak > 54.6821 & rasa > 94.5003	*Dominance*

4.2.3 Evaluasi Performansi

Setelah didapatkan hasil pada tahap klasifikasi, selanjutnya menghitung performansi untuk model prediksi yang telah dibuat. Perhitungan performansi dilakukan dengan percobaan sebanyak 5 kali dengan menginputkan data latih dan data uji kemudian dihitung rata-rata performansi dari tiap percobaan dan menghitung akurasi dengan *Bi-Class* dan *F-Measure*. Berikut adalah hasil performansi dari berbagai macam fitur yang digunakan seperti *Followers*, *Following*, *URL*, *Retweet* dan *Like* (pendekatan perilaku sosial dan linguistik) serta pembagian data set yang berbeda-beda.

Tabel 8 Hasil Akurasi pada Skenario Pendekatan Perilaku Sosial

No.	Data Set	Akurasi *Bi-Class*			
		D	I	S	C
1	70 : 30	96.75	60	46.72	70.23
2	80 : 20	97.5	70.13	57.14	66.67
3	90 : 10	97.56	61.54	47.73	72.34

Dari **Tabel 8** percobaan melalui pendekatan perilaku sosial didapatkan hasil akurasi *Bi Class* Dominance adalah 97.56 dengan ratio perbandingan data latih dan data uji 90 : 10. Hasil dari perhitungan *Bi Class* Influence adalah sebesar 70.13 dengan perbandingan data latih dan data uji 80 : 20. Hasil dari perhitungan *Bi Class* Steadiness adalah sebesar 57.14 dengan perbandingan data latih dan data uji 80 : 20. Hasil dari perhitungan *Bi Class* Compliance adalah sebesar 73.34 dengan perbandingan data latih dan data uji 90 : 10.

Tabel 9 Hasil Perhitungan *F-Measure* pada Perilaku Sosial

No.	Data Set	Precision	Recall	F-Measure
1	70 : 30	0.2549	0.2541	0.2545
2	**80 : 20**	**0.2577206**	**0.2676**	**0.2626**
3	90 : 10	0.2428	0.2458	0.2443

Dari **Tabel 9** percobaan melalui pendekatan perilaku sosial didapatkan hasil dari *F-Measure* pada perilaku sosial sebesar 0.2626 dari data latih dan data uji 80 :20 yang telah dilakukan menggunakan *recall* dan *precision.*

Untuk menghitung performansi pada fitur pendekatan linguistic pada perhitungan *Bi Class* dan *F-Measure*, diperlukan kata-kata terbaik yang dibutuhkan untuk setiap kelas yang ada. Kata-kata terbaik untuk setiap kelas akan terus bertambah seiring dengan jumlah kata yang kita masukkan untuk dihitung nilai performansinya. Berikut merupakan beberapa kata terbaik yang diambil sebagai contoh:

Tabel 10 Hasil Nilai *Bi Class* pada TF-IDF

No.	Data Set	Jumlah Kata	Akurasi *Bi-Class*			
			D	I	S	C
1.	70 : 30	15	96.77	68.55	45.16	70.16
2.	70 : 30	30	96.74	71.54	42.28	69.11
3.	70 : 30	45	95.93	70.73	43.09	68.29
4.	80 : 20	15	96.25	72.25	43.75	70.00
5.	80 : 20	30	96.25	72.50	41.25	70.00
6.	80 : 20	45	96.34	73.17	45.12	70.73
7.	90 : 10	15	97.44	69.23	46.15	74.36
8.	90 : 10	30	97.50	70.00	42.50	70.00
9.	90 : 10	45	97.50	72.50	45.00	70.00

Dari **Tabel 10** percobaan melalui pendekatan linguistik TF-IDF didapatkan hasil akurasi *Bi Class* Dominance adalah 97.50 dengan ratio perbandingan data latih dan data uji 90 : 10 dengan pemilihan 30 dan 45 kata terbaik. Hasil dari perhitungan *Bi Class* Influence adalah sebesar 73.17 dengan perbandingan data latih dan data uji 80 : 20 dari pemilihan 45 kata terbaik. Hasil dari perhitungan *Bi Class* Steadiness adalah sebesar 46.15 dengan perbandingan data latih dan data uji 90 : 10 dengan pemilihan 15 kata terbaik . Hasil dari perhitungan *Bi Class* Compliance adalah sebesar 74.36 dengan perbandingan data latih dan data uji 90 : 10 dengan pemilihan jumlah 15 kata terbaik.

Tabel 11 Hasil Perhitungan *F-Measure* pada TF-IDF

No.	Data Set	Jumlah Kata	Precision	Recall	F- Measure
1.	70 : 30	15	0.2679	0.2910	0.2790
2.	70 : 30	30	0.2472	0.2360	0.2414
3.	70 : 30	45	0.2464	0.2350	0.2405
4.	80 : 20	15	0.2706	0.2974	0.2834
5.	80 : 20	30	0.2458	0.2093	0.2261
6.	80 : 20	45	0.2458	0.2093	0.2261
7.	**90 : 10**	**15**	**0.2580**	**0.3611**	**0.3010**
8.	90 : 10	30	0.2458	0.2093	0.2261
9.	90 : 10	45	0.2458	0.2093	0.2261

Dari **Tabel 11** percobaan melalui pendekatan linguistik didapatkan hasil dari *F-Measure* pada TF-IDF sebesar 0.3010 dari data latih dan data uji 90 :10 yang telah dilakukan menggunakan *recall* dan *precision* dengan jumlah pemilihan kata terbaik sejumlah 15.

Tabel 12 Hasil Nilai *Bi Class* pada TF-RF

No.	Data Set	Jumlah Kata	Akurasi *Bi-Class*			
			D	I	S	C
1.	70 : 30	15	96.74	67.48	44.72	70.73
2.	70 : 30	30	96.72	50.82	53.28	71.31
3.	70 : 30	45	96.77	58.87	47.58	70.97
4.	80 : 20	15	96.34	67.07	41.46	70.73
5.	80 : 20	30	96.34	64.63	42.68	69.51
6.	80 : 20	45	96.29	72.84	43.21	71.60
7.	90 : 10	15	97.50	75.00	45.00	72.50
8.	90 : 10	30	100	61.54	46.15	69.23
9.	90 : 10	45	100	55.26	47.37	71.05

Dari **Tabel 12** percobaan melalui pendekatan linguistik TF-RF didapatkan hasil akurasi *Bi Class* Dominance adalah 100 dengan ratio perbandingan data latih dan data uji 90 : 10 dengan pemilihan 30 dan 45 kata terbaik. Hasil dari perhitungan *Bi Class* Influence adalah sebesar 75.00 dengan perbandingan data latih dan data uji 90 : 10 dari pemilihan 15 kata terbaik. Hasil dari perhitungan *Bi Class* Steadiness adalah sebesar 53.28 dengan perbandingan data latih dan data uji 70 : 30 dengan pemilihan 30 kata terbaik . Hasil dari perhitungan *Bi Class* Compliance adalah sebesar 72.50 dengan perbandingan data latih dan data uji 90 : 10 dengan pemilihan jumlah 15 kata terbaik.

Tabel 13 Hasil Perhitungan *F-Measure* pada TF-RF

No.	Data Set	Jumlah Kata	Precision	Recall	F-Measure
1.	70 : 30	15	0.2572	0.2954	0.2750
2.	70 : 30	30	0.2703	0.3069	0.2874
3.	70 : 30	45	0.2593	0.2992	0.2779
4.	80 : 20	15	0.2401	0.275	0.2563
5.	80 : 20	30	0.2388	0.1625	0.1934
6.	80 : 20	45	0.2647	0.4372	0.3298
7.	90 : 10	15	0.2811	0.4831	0.3554
8.	**90 : 10**	**30**	**0.2412**	**0.1639**	**0.4824**
9.	90 : 10	45	0.2471	0.1754	0.2051

Dari **Tabel 13** percobaan melalui pendekatan linguistik didapatkan hasil dari *F-Measure* pada TF-RF sebesar 0.4824 dari data latih dan data uji 90 :10 yang telah dilakukan menggunakan *recall* dan *precision* dengan pemilihan 30 kata terbaik.

4.3. Hasil Uji

Dari seluruh hasil percobaan dengan menggunakan beberapa fitur dan percobaan penggunaan rasio data set, didapat hasil yang berbeda-beda. Pada pendekatan perilaku sosial dengan fitur-fitur twitter seperti *follower, following, mention, hastag, url , retweet,* dan *like* maka didapatkan nilai hasil akurasi *Bi Class* Dominance adalah 97.56 dengan ratio perbandingan data latih dan data uji 90 : 10. Hasil dari perhitungan *Bi Class* Influence adalah sebesar 70.13 dengan perbandingan data latih dan data uji 80 : 20. Hasil dari perhitungan *Bi Class* Steadiness adalah sebesar 57.14 dengan perbandingan data latih dan data uji 80 : 20. Hasil dari perhitungan *Bi Class* Compliance adalah sebesar 73.34 dengan perbandingan data latih dan data uji 90 : 10.

Pada perhitungan percobaan melalui pendekatan perilaku sosial dengan fitur-fitur twitter *follower, following, mention, hastag, url , retweet,* dan *like* didapatkan hasil perhitungan *F-Measure* sebesar 0.2626 dari data latih dan data uji 80 :20.

Pada perhitungan percobaan melalui pendekatan linguistik TF-IDF maka didapatkan hasil dari tiap-tiap kelas nya dengan perhitungan akurasi *Bi Class,* Dominance adalah 97.50 dengan ratio perbandingan data latih dan data uji 90 : 10 dengan pemilihan 30 dan 45 kata terbaik. Hasil dari perhitungan *Bi Class* Influence adalah sebesar 73.17 dengan perbandingan data latih dan data uji 80 : 20 dari pemilihan 45 kata terbaik. Hasil dari perhitungan *Bi Class* Steadiness adalah sebesar 46.15 dengan perbandingan data latih dan data uji 90 : 10 dengan pemilihan 15 kata terbaik . Hasil dari perhitungan *Bi Class* Compliance adalah sebesar 74.36 dengan perbandingan data latih dan data uji 90 : 10 dengan pemilihan jumlah 15 kata terbaik.

Pada perhitungan percobaan melalui pendekatan linguistik TF-RF maka didapatkan hasil dari tiap-tiap kelas nya dengan perhitungan akurasi *Bi Class,* Dominance adalah 100 dengan ratio perbandingan data latih dan data uji 90 : 10 dengan pemilihan 30 dan 45 kata terbaik. Hasil dari perhitungan *Bi Class* Influence adalah

sebesar 75.00 dengan perbandingan data latih dan data uji 90 : 10 dari pemilihan 15 kata terbaik. Hasil dari perhitungan *Bi Class* Steadiness adalah sebesar 53.28 dengan perbandingan data latih dan data uji 70 : 30 dengan pemilihan 30 kata terbaik . Hasil dari perhitungan *Bi Class* Compliance adalah sebesar 72.50 dengan perbandingan data latih dan data uji 90 : 10 dengan pemilihan jumlah 15 kata terbaik.

Hasil perhitungan *F-Measure* pada pendekatan linguistik TF-IDF sebesar 0.3010 dari data latih dan data uji 90 :10 yang telah dilakukan menggunakan *recall* dan *precision* dengan jumlah pemilihan kata terbaik sejumlah 15. Hasil perhitungan *F-Measure* pada pendekatan linguistik TF-RF sebesar 0.4824 dari data latih dan data uji 90 :10 yang telah dilakukan menggunakan *recall* dan *precision* dengan pemilihan 30 kata terbaik.

5. Kesimpulan

Hasil pengujian pada penelitian ini, analisa kepribadian DISC dari pengguna twitter dengan menggunakan metode Decision Tree C.45 dengan pembobotan TF-IDF dan TF-RF bertujuan untuk mendapatkan nilai performansi yang tinggi. Hal ini terlihat dari hasil uji yang telah dilakukan, dimana sistem mampu mengklasifikasi data uji ke dalam empat kategori yaitu *Dominance*, *Influence*, *Steadiness*, dan *Compliance.* Hasil dari penelitian yang telah dilakukan seperti perhitungan nilai *Bi Class* menyatakan bahwa masing- masing kelas mendapatkan hasil prediksi dengan pendekatan perilaku sosial dan pendekatan linguistik. Hasil penelitian dengan perhitungan *F-Measure* dengan percobaan melalui pendekatan perilaku sosial dengan fitur-fitur twitter *follower, following, mention, hastag, url , retweet,* dan *like* didapatkan hasil perhitungan *F-Measure* sebesar 0.2626 dari data latih dan data uji 80 :20. Dan hasil akurasi pendekatan lingustik menggunakan pembobotan TF-IDF sebesar 0.3010 dari data latih dan data uji 90 :10 yang telah dilakukan menggunakan *recall* dan *precision* dengan jumlah pemilihan kata terbaik sejumlah 15. Hasil akurasi pendekatan lingustik menggunakan pembobotan TF-RF sebesar 0.4824 dari data latih dan data uji 90 :10 yang telah dilakukan menggunakan *recall* dan *precision* dengan pemilihan 30 kata terbaik.

Pada penelitian ini dapat ditarik kesimpulan meskipun sudah mendapatkan data dari hasil psikotes mahasisawa Telkom University 2013 berdasarkan data tersebut masih mendapatkan ketimpangan data dimana jumlah data lebih banyak terdapat pada kelas *Steadiness* dan *Compliance* sehingga membuat nilai akurasi pada model prediksi yang dibangun kecil karena tidak seimbang dari tiap kelasnya.

Saran untuk penelitian selanjutnya, untuk mencari algoritma tambahan atau referensi mengenai penanganan dalam ketimpangan suatu data pada kelas tertentu agar keputusan pada model prediksi tidak cenderung kepada kelas data yang dominan dan dilakukan uji coba menggunakan data yang lebih banyak agar tingkat akurasi aplikasi dapat ditingkatkan.

Daftar Pustaka

[1] A. V. Member and G. Mohammadi, “Vinciarelli, A., and Mohammadi, G. (2014),” vol. 5, no. December, pp. 273–291, 2014.

[2] L. N. Rani, “Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit,” *J. KomTekInfo Fak. Ilmu Komput.*, vol. 2, no. 2, pp. 33–38, 2015.

[3] A. . Fallis, “Kajian pustaka,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

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

[5] F. Dwi Meliani Achmad, Budanis, Slamat, “Klasifikasi Data Karyawan Untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree,” *J. IPTEK*, vol. 16, no. 1, pp. 18–23, 2012.

[6] et al Riwayati, “Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2014 Yogyakarta, 15 November 2014 ISSN: 1979-911X,” *Snast*, vol. 3, no. November, pp. 211–216, 2014. [7] Inscape Partners, “The DiSC model,” *Inscape Publishing.*, pp. 1–16, 2006.

[7] D. Lhaksmana, KM; Nhita, Fhira & Anggraini, “Klasifikasi Kepribadian Berdasarkan Status Facebook Menggunakan Metode Backpropagation,” *e-Proceeding Eng.*, vol. 4, no. 3, pp. 5174–5183, 2017.

[8] H. Jantan, A. Razak Hamdan, and Z. Ali Othman, “Human Talent Prediction in HRM using C4 . 5 Classification Algorithm,” *Int. J. Comput. Sci. Eng.*, vol. 02, no. 08, pp. 2526–2534, 2010.

[9] K. Tampubolon, H. Saragih, B. Reza, K. Epicentrum, A. Asosiasi, and A. Apriori, “Implementasi Data Mining Algoritma Apriori Pada Sistem Persediaan Alat-alat Kesehatan,” *Implementasi Data Min. Algoritm. Apriori Pada Sist. Persediaan Alat-alat Kesehat.*, vol. Volume : I, no. Majalah Ilmiah Informasi dan Teknologi Ilmiah (INTI) ISSN : 2339-210X, pp. 93–106, 2013.

[10] J. C. García-López, J. M. Pinos-Rodríguez, I. A. García-Galicia, G. B. Galicia-Juárez, M. L. Gorostiola-Herrera, and M. A. Camacho-Escobar, “Efecto Del Uso De Una Enzima Y Sistema De Alimentación Sobre Productividad En Pavos,” *Arch. Zootec.*, vol. 60, no. 230, pp. 297–300, 2011.

[11] O. Korukcu and K. Kukulu, “The challenges and opportunities of social media in health,” *Turkish Online J. Educ. Technol.*, vol. 2016, no. DecemberSpecialIssue, pp. 743–745, 2016.

[12] M. M. Degree, C. Science, and A. C. Lecture, “Data Mining : Concepts and,” vol. 05, p. 703, 2012.

[13] C. K. E. Goni, H. Opod, and L. David, "Gambaran Kepribadian Berdasarkan Tes DISC Mahasiswa Fakultas Kedokteran Universitas Sam Ratulangi Manado," *J. E-Biomedik*, vol. 4, no. 2, 2016.
[14] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti," *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017.
[15] D. C. S. Attainer and D. I. Concluder, "25 Personality Styles based on DISC SIC, SCI -Advocate."
[16] A. A. Maarif, "Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah," *Penerapan Algoritm. TF-IDF untuk Pencarian Karya Ilm.*, no. 5, p. 4, 2015.
[17] A. T. Damanik and M. L. Khodra, "Prediksi Kepribadian Big 5 Pengguna Twitter dengan Support Vector Regression," *J. Cybermatika*, vol. 3, no. 1 (3), pp. 14–22, 2015.
[18] Agnes, theresia dan leylia.2015. prediksi kepribadian big 5 pengguna Twitter dengan support vector regression, jurnal cybermatika vol.3 no.1.
[19] B. Verhoeven, W. Daelemans, and B. Plank, "Twisty: a Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling," *Proc. 10th Lang. Resour. Eval. Conf. (LREC 2016)*, pp. 1632–1637, 2016
[20] Fadillah, Sarah. (2013). "Implementasi Data Mining Untuk Pengenalan Karakteristik Transaksi Customer Dengan Menggunakan Algoritma C4.5." Pelita Informatika Budi Darma, Vol. 5, No. 3. 2301-9425.
[21] Ginting, Selvia Lorena Br., Zarman, Wendi, dan Hamidah, Ida. (2014) "Analisis dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik." Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNASTI). 1979- 911X. [1]
[22] A. M. Buckingham, D. O. Clifton, and D. Ph, "Now , Discover Your Strengths," vol. 5, no. 2, 2003.
[23] J. H. Reynierse, D. Ackerman, A. a Fink, and J. B. Harker, "The Effects of Personality and Management Role on Perceived Values in Business Settings," *Int. J. ValueBased Manag.*, vol. 13, no. 1992, pp. 1–13, 2000.
[24] A. D. Davies and A. D. Davies, "Freedom Movement Published by : Wiley on behalf of The Royal Geographical Society (with the Institute of Ethnography , space and politics : interrogating the process of protest in the Tibetan Freedom Movement," vol. 41, no. 1, pp. 19–25, 2017.
[25] Badri, 2011:132
[26] Wu Harry dan Salton Gerald. *A Comparison of Search Term Weigthing: Term Relevance vs Inverse Document Frequency.*1981.
[27] Wu Haibing dan Gu Xiaodong, *Reducing Over-Weighting in Supervised Term Weighting for Sentiment Analysis.*2014.