Prediksi Kepribadian DISC dengan *K-Nearest Neighbors Algorithm* (KNN) Menggunakan Pembobotan TF-IDF dan TF-*Chi Square*

Nur Ihsan Putra Munggaran[1], Erwin Budi Setiawan[2]
[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]putraihsan@students.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id,

Abstrak

Media sosial berkembang pesat pada saat ini. Salah satu media sosial yang berkembang dengan sangat pesat adalah twitter. Twitter adalah media sosial yang di dalamnya berisikan informasi seperti biografi seseorang dan *tweet* atau cuitan dari penggunanya. Oleh karena informasi yang kita dapatkan di twitter bisa dimanfaatkan untuk menggambarkan kepribadian seseorang. Ada banyak metode yang bisa digunakan untuk mengetahui kepribadian seperti Big 5, MBTI dan DISC. Dalam penelitian ini penulis menggunakan metode DISC (*Dominance Influence Steadiness Conscientiousness*) karena metode ini masih sangat sedikit digunakan untuk penelitian, dan penulis menggunakan metode pengklasifikasian dari data *mining* dengan metode pengklasifikasian *K- Nearest Neighbors Algorithm* (KNN). Fitur *linguistic* yang akan digunakan dibagi menjadi dua bagian yaitu fitur katagori kata dari *corpus* yang dibangun secara manual dan fitur yang didapatkan langsung dari data twitter menggunakan twitter *apps.* Penelitian ini akan sangat berguna untuk pemilihan sumber daya manusia karena bisa menghemat biaya dan tenaga yang dikeluarkan, dengan menggunakan aplikasi ini bisa menentukan kepribadian seseorang hanya dengan menggunakan media sosial twitter. Metode pembobotan yang digunakan dalam penelitian ini adalah TF-IDF dan TF-*Chi Square* yang berguna untuk mengukur bobot setiap kata pada sebuah *tweet.* Dari hasil percobaan didapatkan akurasi terbaik sebesar 40.60% pada perbandingan data latih dan data uji sebesar 60:40 dan pendekatan linguistik dengan menggunakan skenario pendakatan perilaku sosial dengan pemilihan nilai k sebesar 61.

Kata Kunci : DISC, KNN, TF-IDF, TF-*Chi Square, Twitter*

Abstract

Social media is growing rapidly at this time. One of the fastest growing social media is Twitter. Twitter is social media which contains information such as someone's biography and tweets or tweets from users. Because the information we get on twitter can be used to describe a person's personality. There are many methods that can be used to find out personalities such as Big 5, MBTI and DISC. In this study the author uses the DISC (Dominance Influence Steadiness Conscientiousness) method because this method is still very little used for research, and the author uses the classification method of data mining with classification methodsK-Nearest Neighbors Algorithm (KNN). The linguistic feature that will be used is divided into two parts, namely the word category feature from the manually built corpus and features obtained directly from twitter data using twitter apps. This research will be very useful for the selection of human resources because it can save costs and labor spent, using this application can determine one's personality using only social media twitter. The weighting method used in this study is TF-IDF and TF- Chi Square which is useful for measuring the weight of each word in a tweet . From the results of the experiments obtained accuracy the best is 40. 60 % in the comparison of training data and test data at 60:40 and the linguistic approach using a social behavior approach scenario with the selection of a k value of 61.

Keywords : DISC, KNN, TF-IDF, TF-Chi Square, Twitter

1. Pendahuluan

Perkembangan teknologi pada saat ini sudah semakin pesat dengan adanya media sosial seperti Facebook, Twitter, Instagram dan sejenisnya. Pengguna sosial media pada January 2005 sudah mencapai 115 juta member[1]. Salah satu media sosial twitter. Twitter adalah sosial media yang berbasis *microblogging* yang peluncurannya pada tanggal 13 Juli 2006[2]. Pengguna twitter saat pada tahun 2009 sudah mencapai lebih dari 41 juta orang dan terus meningkat hingga saat ini[3]. *Tweet* yang yang di-*posting* seseorang itu bisa menggambarkan perasaan dan kepribadian dari orang tersebut, oleh karena itu dalam penulisan penelitian ini penulis ingin mengambil informasi *tweet* dari twitter sebagai bahan acuan kepribadian seseorang.

Untuk menilai kepribadian seseorang biasanya bisa menggunakan beberapa metode seperti kuisioner, wawancara dan juga menggunakan jasa psikolog tetapi membutuhkan banyak waktu dan dana yang lebih. Saat ini sudah banyak metode yang dapat digunakan untuk menilai kepribadian. Salah satu metode yang bisa digunakan ialah kepribadian DISC.

Kepribadian DISC dikemukakan oleh seorang 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)*[4]. Kepribadian DISC merupakan salah satu alat penilai kepribadian seseorang tetapi belum banyak digunakan untuk penelitian oleh

karena itu DISC merupakan alternatif yang baru dalam melakukan penelitian[4]. Dalam penelitian sebelumnya[4] metode kepribadian DISC mengklasifikasikan kepribadian dengan data yang pengambilannya menggunakan test kepada setiap mahassiswa. Oleh karena itu penulis ingin melakukan penelitian penilaian kepribadian dengan menggunakan data *tweet* di twitter.

Algoritma yang terdapat pada data *mining* dapat digunakan sebagai salah satu metode dalam penelitian ini. Data *mining* bisa di sebut juga pencarian pengetahuan dari data yang dimana data *mining* akan mengesktrak secara otomatis pola atau pengetahuan yang menarik, tersembunyi, tidak diketahui sebelumnya, dari data dalam jumlah yang sangat besar[5]. Data mining merupakan ilmu yang relatif baru, dan sampai sekarang orang masih memperdebatkan untuk menempatkan data mining di bidang ilmu mana karena data mining menyangkut database, kecerdasan buatan, statistik[6]. Salah satu algoritma pengklasifikasian data yang terdapat dalam data mining adalah *K- Nearest Neighbors Algorithm* (KNN)[7].

KNN adalah algoritma yang digunakan untuk melakukan klasifikasi terhadap suatu objek, berdasarkan K buah data latih yang jaraknya paling dekat dengan objek tersebut[8].

Dalam penelitian ini, penulis menggunakan pembobotan *Term Frequency Inverse Document Frequency* (TF-IDF) dan TF-*Chi Square*. TF-IDF yang merupakan metode yang digunakan dalam melakukan pembobotan terhadap kemunculan kata dalam suatu dokumen[7]. Uji TF-*Chi Square* berguna untuk menguji hubungan dua buah variabel nominal dan mengukur kuatnya hubungan antara variabel yang satu dengan variabel nominal lainnya. Oleh Karena itu, pada penelitian ini akan melihat seberapa akurat penggunaan algoritma KNN dalam pengklasifikasian kepribadian pengguna twitter serta seberapa besar perbedaan antara dua metode pembobotan yang digunakan.

2. Studi Terkait

2.1 Kepribadian DISC

Kepribadian adalah keseluruhan cara seorang individu bereaksi dan berinteraksi dengan individu lain paling sering dideskripsikan dalam istilah sifat yang bisa diukur yang ditunjukkan oleh seseorang[4]. Terdapat berbagai macam teori untuk mencari kepribadian seseorang seperti *Big* 5, MBTI, dan DISC. Pada penelitian kali ini penulis akan menggunakan teori DISC. Teori DISC ditemukan oleh ahli psikolog asal Amerika yang bernama William Moulton Marston pada tahun 1928. Ia berteori bahwa perilaku emosi bisa dikatagorikan menjadi 4 tipe perilaku individu ketika berinteraksi dengan lingkungan yaitu *Dominance (D)*, *Influence (I)*, *Steadiness (S)*, dan *Compliance (C)*[4].

Dominance (D) adalah orang yang bertipe tegas, abisius, independen, menyukai persaingan, penerima tantangan, cepat dalam mengambil keputusan, penuntut, tidak sabar, dan tidak menyukai hal yang rutin. *Influence (I)* adalah orang yang bertipe ramah, senang bergaul, suka menghibur orang lain, antusias, optimis, motivator, kurang memerhatikan detail, banyak bicara, mudah lupa, dan seringkali bereaksi berlebihan terhadap sesuatu. *Steadiness (S)* adalah orang yang bertipe sabar, gigih, jujur, akomodatif, loyal, tidak terlalu menuntut, ingin menolong orang lain, tidak suka dengan perubahan, kurang antusias, kurang tegas, cenderung menghindar dari konflik, dan sulit menyusun prioritas. *Compiliance (C)* adalah orang yang bertipe teliti, terstruktur, berhati-hati dalam membuat keputusan, kritis dalam membuat keputusan, kritis dalam menganalisa, kerja sendiri maupun kelompok, patah terhadap atasan atau pimpinan, kurang fleksibel, defensif ketika di kritik, terlalu mengikuti aturan dan lamban saat menyelesaikan tugas[4].

Tabel 1. Sistem DISC

	Ciri Umum	Nilai Dalam Team	Kemungkinan Kelemahan	Ketakutan Terbesar
D	Langsung; Tegas; rasa ego yang tinggi; *Problem solver; Risk taker; Self-starter*	*Bottom-line organizer;* Menghargai waktu; Menantang status; Inovatif	Melanggar kewenangan; Sikap argumentif; Menolak rutinitas; Cenderung mengerjakan banyak hal dalam waktu yang bersamaan	Dimanfaatkan orang lain
I	Antusias; Percaya; Optimis; Persuasif; Bicara aktif; Impulsif; Emosional	*Problem Solver* yang Kreatif; Penggugah semangat yang baik; Memotifasi orang lain; Selera humor yang positif;	Mencari popularitas dari hasil kerja nyata; Kurang perhatikan *detail*; Terlalu menggunakan	Penolakan

		Menengahi konflik; Pembawa damai	bahasa tubuh; Mendengar hanya bagian kesukaannya	
S	Pendengar yang baik; *Team player*; Possesive; Stabil; Dapat diprediksi; Memahami orang lain; Bersahabat	Dapat di percaya dan diandalkan; Anggota team yang loyal; Taat akan otoritas; Pendengar yang baik; Sabar dan berempati; Mendamaikan konflik	Menolak perubahan; Butuh waktu lama untuk berubah; Menyimpan dendam; Sensitif pada kritik; Sulit menentukan prioritas	Kehilangan rasa aman
C	Akurat; Analitis; Cermat; Hati-hati; *Fact-Finder;* Presisi tinggi; Standar kinerja tinggi; Sistematis	Rajin dan hati-hati; Tuntas dalam kegiatan; Menggambarkan situasi; Mengumpulkan, mengkritisi dan menguji informasi	Membutuhkan batasan yang jelas; Terikat pada prosuder dan metoda; Sangat detail; Tidak mengungkapkan perasaan; Cenderung menerima dari pada argumentasi	Kritik

2.2 Penggunaan Fitur

Seperti yang sudah dijelaskan, pada penelitian ini penulis melakukan analisis yaitu 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 [1] 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 tanda baca adalah banyaknya simbol sebuah dari sebuah kata yang ingin diungkapkan oleh pengguna, tanda baca yang dihitung adalah tanda tanya (?) dan tanda seru (!).
d. 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 [16] disimpan untuk dimasukan kedalam kamus pada *database*. Total emoji yang didapat berjumlah 2.552 karakter.
e. Rata-rata kata adalah jumlah dari kata yang dituliskan pengguna *twitter* dibagi dengan jumlah *tweet* yang berhasil di *crawling*.
f. Jumlah huruf besar adalah banyaknya huruf kapital yang digunakan pengguna saat menulis *tweet.*
g. Jumlah karakter adalah dari susunan huruf atau simbol-simbol yang menyusun sebuah *tweet.*
h. Rata-rata karakter adalah rata-rata jumlah dari karakter yang dituliskan pengguna di *tweet* di bagi dengan jumlah *tweet* yang berhasil di *crawling.*

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 *Chi-Square*.

2.3 Pre-Processing

Preprocessing data adalah tahapan untuk mengubah data acuan menjadi lebih terstruktur. Berikut adalah beberapa proses *preprocessing* data yang dilakukan terhadap *tweet* .

a. *Case Folding*, proses mengubah semua karakter dalam dokumen ke dalam kasus yang sama, baik semua huruf besar atau huruf kecil, untuk mempercepat perbandingan selama proses pengindeksan. Biasanya dalam berbagai kasus, semua huruf akan dijadikan huruf kecil.
b. *Tokenization*, mengubah kalimat menjadi kumpulan satu kata. *Tokenization* seakan-akan hanya perlu membagi *string* (kata) pada setiap ruang, akan tetapi tidak berhenti sampai disitu. Beberapa singkatan seperti ‘dll’ dan ‘e.g.’ biasanya ditulis dengan tanda baca, jadi tanda baca tidak selamanya mengakhiri sebuah kalimat. Tanda hubung, digit, dan lainnya juga ditangani.
c. *Filtering*, menghilangkan *stop word*. *Stop word* adalah kata umum yang memiliki sedikit atau tanpa makna, tetapi diperlukan dalam struktur bahasa gramatikal. Beberapa contoh *stop word* pada bahasa Indonesia diantaranya ‘ýang’, ‘di’, dan ‘ke’.
d. *Stemming*, mengembalikan kata ke dalam bentuk dasar (kata dasar) dengan menghilangkan aditif yang ada. Salah satu contohnya ialah kata ‘*connection*’ dapat direduksi menjadi kata ‘*connect*’. Algoritma Stemming yang bisa digunakan adalah Porter Stemmer untuk bahasa Inggris dan Nazief-Andriani untuk bahasa Indonesia.

2.4 Pembobotan

Pembobotan merupakan suatu proses yang melibatkan semua faktor secara bersamaan dengan cara memberi bobot kepada masing-masing faktor tersebut. Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan istilah dalam dokumen karena dari situ bisa dipercaya bahwa frekuensi kemunculan istilah merupakan petunjuk sejauh mana istilah tersebut dalam mewakili isi dari dokumen[11]. Ada berbagai cara untuk menyelesaikan metode pembobotan namun pada penelitian ini penulis menggunakan TF-IDF dan TF-*Chi Square*.

2.4.1 *Term Frequency Inverse Document Frequency*

Term Frequency Inverse Document Frequency(TF-IDF) merupakan metode yang digunakan dalam melakukan pembobotan terhadap kemunculan kata dalam suatu dokumen[7]. Pembobotan TF-IDF dapat dirumuskan sebagai berikut:

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

Keterangan:

$tfidf_t$ = Bobot total dari data ke-t
$f_{t,d}$ = Kemunculan kata ke-t dalam dokumen ke-d
N = Total dokumen
df_t = banyaknya dokumen yang mengandung kata ke-t

2.4.2 *Term Frequency Chi-Square*

TF *Chi-Square* berguna untuk menguji hubungan atau pengaruh dua buah variabel nominal dan mengukur kuatnya hubungan antara variabel yang satu dengan variabel nominal lainnya [14]. TF *Chi–square* mempertimbangkan bobot untuk *term* yang tidak muncul dalam dokumen dan *term* yang muncul didalam dokumen, dan dapat ditulis dengan persamaan [15]:

$$TF \times X^2(t, c) = TF\ (d, t) \times \frac{\mathrm{N\ (AD-BC)2}}{(A+B)(C+D)(A+C)(B+D)} \quad (2)$$

Dimana:

TF (d, t) : Frekuensi kemunculan kata t pada dokumen *tweet* d
N : Total Keseluruhan dokumen
A : Banyaknya dokumen *tweet* dalam kelas c yang mengandung *term* t
B : banyaknya dokumen *tweet* yang tidak terdapat pada kelas *c* namun memuat *term t*
C : banyakanya dokumen *tweet* yang terdapat pada kelas *c* namun tidak memuat *term t*
D : banyaknya dokumen *tweet* yang bukan merupakan kelas c dan tidak memuat *term t*

c : Kelas kategori
t : Kata

2.5 *K-Nearest Neighbor* (KNN)

Klasifikasi merupakan sebuah metode untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep antar tiap kelas data, dengan tujuan dapat memperkirakan kelas yang labelnya tidak diketahui. *k-Nearest Neighbor (*KNN*)* adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (*train dataset*), diambil dari k tetangga terdekatnya (*nearest neighbors*), dengan k merupakan banyaknya tetangga. Salah satu metode yang menerapkan algoritma *supervised.* Perbedaan antara *supervised learning* dengan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang suda ada, dan sedangkan *unsupervised learning*, data belum memiliki pola apapun. Berdasarkan kategori pada algoritma KNN, dimana kelas yang paling banyak muncul nantinya akan menjadi kelas dari hasil klasifikasi. Berikut merupakan langkah-langkah algoritma KNN :

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

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

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

Keterangan
D :jarak antara titik
x :*data training*
y :*data testing*

2.6 Evaluasi Performansi

Karena dalam melakukan klasifikasi yang didapat tidak akan bekerja 100% benar oleh karena itu terdapat sejumlah ukuran yang dapat mengevaluasi model klasifikasi diantaranya *recall* dan *precision.* Berikut adalah tabel kontingensi yang berfungsi menganalisis kualitas *classifier* dalam mengenali tuple-tuple dari kelas yang tersedia.

Tabel 2 Tabel Kotingensi untuk Prediksi dan Aktual

Kategori	**Kelas Prediksi**		
Kelas Aktual		**Ya**	**Tidak**
	Ya	TP	FN
	Tidak	FP	TN

Dimana:
TP (Benar Positif) : Kelas yang diprediksi *yes*, dan faktanya adalah *yes.*
TN (Benar Negatif) : Kelas yang diprediksi *no*, dan faktanya adalah *no.*
FP (Salah Positif) : Kelas yang diprediksi *yes*, dan faktanya adalah *no.*
FN (Salah Negatif) : Kelas yang diprediksi *no*, dan faktanya adalah *yes*

1. Akurasi
   Akurasi merupakan parameter evaluasi terhadap sistem yang dibangun dalam penelitian penelitian ini. Berikut adalah rumus akurasi[7]:

$$Akurasi = \frac{TP+TN}{(TP+FP+TN+FN)} \qquad (4)$$

2. *Precision*

Precision adalah jumlah user yang dengan benar diklasifikasi dalam sebuah kelas dibagi dengan jumlah total *user* yang diklasifikasikan dalam kelas tersebut[7]. Berikut adalah rumus *precision:*

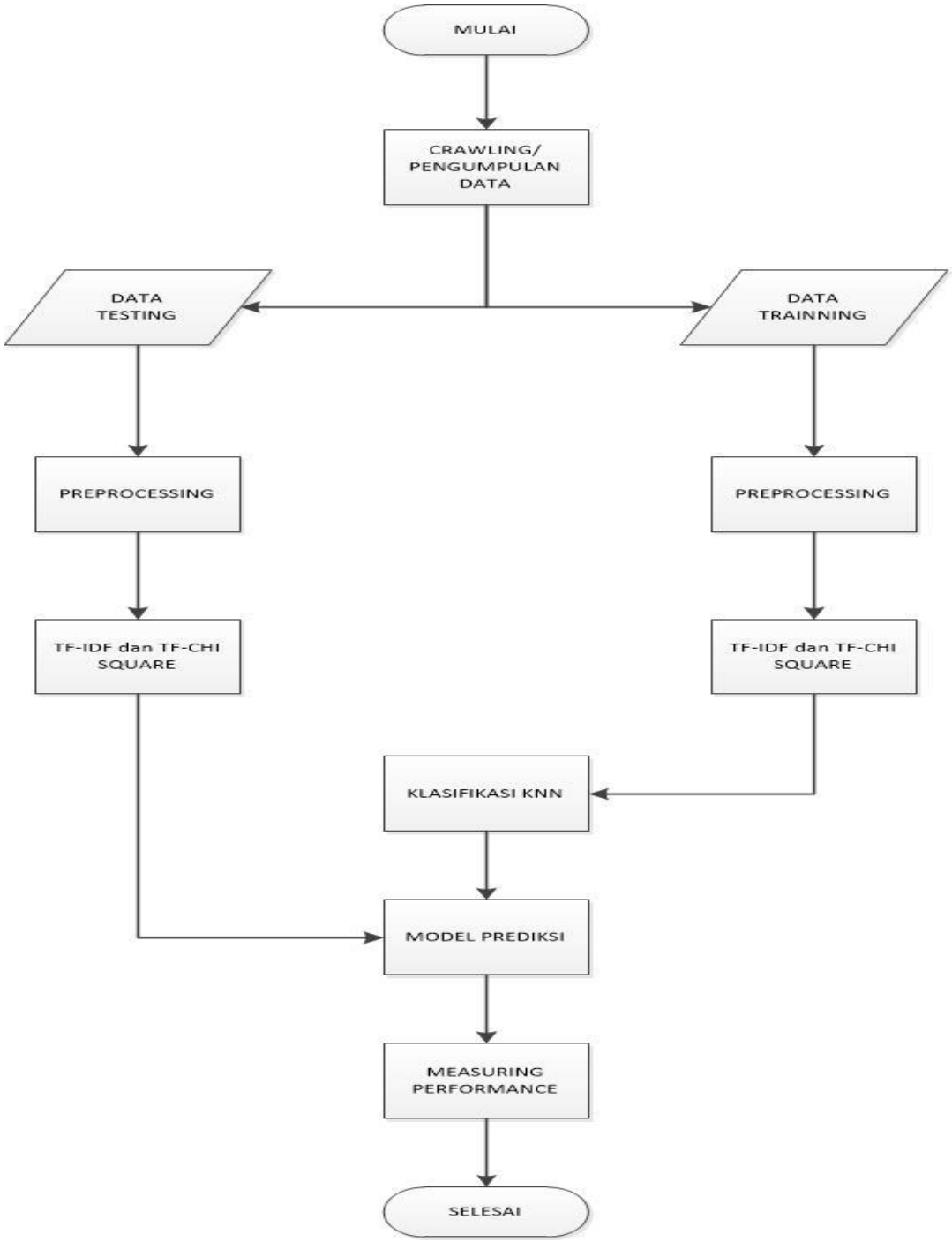
$$Precision = \frac{FP}{(FN+FP)} \quad (5)$$

3. Recall
   Recall adalah jumlah *user* yang dengan benar diklasifikasikan dalam sebuah kelas dibagi dengan jumlah total *user* dalam kelas tersebut[7]. Berikut adalah rumus *recall:*

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

3. Sistem yang Dibangun

Rangkaian proses dalam sistem prediksi kepribadian dengan KNN akan dibangun seperti Gambar 1.



Gambar 1. Sistem Prediksi Kepribadian dengan KNN

1. *Crawling Data*
   Data yang digunakan dalam penelitian ini adahal hasil *tweet* dan beberapa atribut dari fitur yang akan dianalisis. Setiap *crawling* akan melakukan pengambilan tweet sebanyak 3200 *tweet*. Jumlah akun yang dapat *dicrawling* adalah 454 akun, akun tersebut diambil dari data mahasiswa telkom yang melakukan tes psikotes pada tahun 2013 dan untuk total *tweet* yang dapat dikumpulkan sebanyak 1.000.489. Data *crawling* tersebut akan dijadikan data *training* dan 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 (60:40), (70:30), (80:20).

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 *Chi-Square*
   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. Hasil dari pembobotan berupa banyak kata yang telah diberi bobot.

5. Klasifikasi KNN(*K-Nearest Neighbor)*
   Pada proses ini data yang telah di *preprocessing* akan masuk ke tahap klasifikasi. Kemudian data latih akan diinput dan dihitung berdasarkan proses KNN. Output dari proses ini adalah model prediksi yang akan di ajukan nilai performansinya.

6. Evaluasi Performansi
   Evaluasi performansi ialah sebuah tahapan terakhir untuk melihat seberapa besar akurasi, *precision*, dan *recall* yang didapatkan untuk mengukur performansi sistem yang telah dibuat.

7. Pengaruh Parameter Nilai k
   Pada proses ini bertujuan untuk melihat pengaruh nilai k terhadap proses performansi dan dilihat k mana yang mempunyai akurasi yang cukup tinggi

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

Data set didapatkan dari hasil psikotes mahasiswa Telkom 2013, dari lebih kurang 1600 peserta crawling yang didapat hanya 454 dikarenakan ada mahasiswa yang tidak memliki akun *twitter*, akunya sudah tidak ada dan ada juga yang akunnya di kunci. Untuk perincian kelasnya dapat dilihat pada gambar 2.

Setiap mahasiswa yang mengikuti tes psikotes akan mendapatkan hasil berupa kepribadian yang paling dominan sampai yang paling tidak dominan. Berikut ini adalah tabel beberapa contoh dari hasil tes psikotes mahasiswa Telkom tahun 2013.

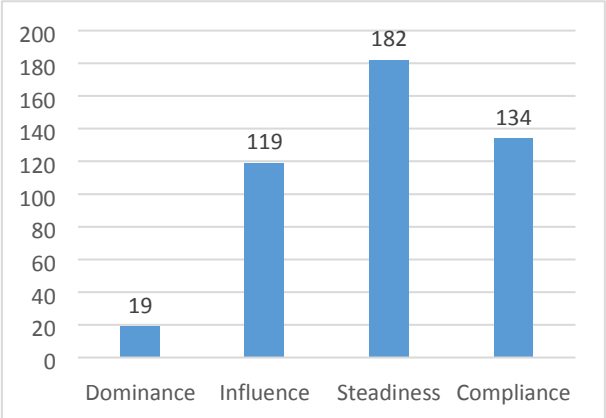
Tabel 4 Data Hasil Psikotes Mahasiswa Universitas Telkom 2013

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	
…	…	…	…

Dari hasil tabel diatas didapatkan kepribadian mahasiswa dari yang paling dominan hingga tidak terlalu dominan melalui huruf pada tabel yang paling pertama. Dari data yang didapat akan diambil yang paling dominan dan dijadikan sebagai kelas. Dalam penelitian ini peneliti hanya memakai 4 katagori oleh karena itu hanya mengambil satu kelas yang paling dominan. Kelas pertama adalah kepribadian D (*Dominance)*, kelas kedua merupakan kepribadian I (*Influance*), kelas ketiga S (*Steadiness*), dan kelas ke empat adalah kepribadian C (*Compliance*).

Tabel 5 Data Hasil Crawling Akun Twitter Peserta Psikotes

ID	Nama	Akun	Follower	Following	Retweet	Hashtag	…	Label
3	Fauzan Abdurrahman	siojanpaujan	35	190	92	151	…	**Compliance**
4	Gandung Tarispranoto	Gandungtaris	4	0	0	0	…	**Steadiness**
…	…	…	…	…	…	…	…	**…**
454	Umar Fatih	Umaaarf	260	251	617	283	…	**Influence**



Gambar 2 List Label yang Telah Disederhanakan Menjadi 4 Kategori

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 bagian ini semua *tweet* akan di kumpul dalam satu bagian per akun yang nantinya akan di lakukan *preprocessing* untuk mengubah data menjadi terstruktur.

Tabel 6 Contoh Hasil Preprocessing

Akun: siojanpaujan	**Sebelum Preprocessing**	"pengen pulangggg keeeruumaaahhh!!" "capE sama KAMU gini terus @_@" "kehidupan baru dengan status yang berbeda sama sekali nggA ADA bedanya -__-" "tersinggung sama sikap mu.."@Denpasr"
	Setelah Preprocessing	sama sikap sama sama hidup baru beda sama

Tahap selanjutnya adalah pembobotan semua kata hasil *preprocessing* akan diberi bobot dengan menggunakan perthitungan TF-IDF dan TF-*Chi Square.*

Tabel 7 Contoh Hasil Pembobotan TF-IDF dan TF Chi-Square

Akun	Kata						
	lama	**Rasa**	**Harap**	**Bias**	**hajar**	**Ingin**	**Sikap**
TF 1	20	25	4	35	6	4	1
TF 2	1	0	0	0	0	0	0
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**
IDF	0.9548	0.7892	1.3332	0.4856	1.078	1.2044	3.974
CHI	1.3156	2.4568	2.1819	1.1334	2.2458	3.0279	2.1513
TF_IDF 1	19.096	19.73	5.3328	16.996	6.468	4.8176	3.974
TF_IDF 2	0.9548	0	0	0	0	0	0
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**
TF_CHI 1	26.312	61.42	8.7276	39.669	13.4748	12.1116	2.1513
TF_CHI 2	1.3156	0	0	0	0	0	0
TF_CHI 3	71.0424	63.8768	13.0914	119.007	24.7038	36.3348	2.1513

4.2.2 KNN(*K-Nearest Neighbor)*

Selanjutnya adalah pengerjaan metode KNN, data yang sudah didapat dari hasil pembobotan dipisahkan menjadi dua, yaitu data *training* dan data *testing* yang kemudian dilatih, untuk mendapatkan kelas prediksinya dan langsung dimasukan kedalam sistem yang sudah dibuat.

Tabel 8 Contoh Hasil Prediksi KNN

ID	Akun	Kelas Asli	kelas Prediksi
1	Gandungtaris	S	S
2	Juanalkautsar	C	S
3	chiiririn	I	S
4	Fia_Utari	I	I
5	denmasniko	S	S

4.2.3 Pengukuran Performansi

Setelah hasil klasifikasi didapatkan lalu akan dilanjutkan evaluasi performansi untuk model prediksi yang telah dibuat, dengan menghitung nilai akurasi, *precision*, dan *recall.* Berikut adalah hasil performansi berbagai macam fitur yang digunakan (pendekatan perilaku social dan linguistik) serta pembagian data set yang berbeda-beda. Dibawah ini adalah tabel pengujian dengan melihat pengaruh nilai K terhadap nilai akurasi.

Tabel 9 Pengaruh Nilai K

No	Data Set	K	Atribut											Akurasi
			FR	**FG**	**MU**	**URL**	**M**	**RT**	**#**	**HB**	**TB**	**EM**	**KT**	
1	60:40	20	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	35,15%
2	60:40	25	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	40,60%
3	60:40	30	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	40,60%
4	60:40	35	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	38,18%

5	60:40	40	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	40,60%
6	**60:40**	**45**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**41,21%**
7	60:40	50	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	40,60%
8	60:40	60	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	39,33%
9	70:30	20	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	33.57%
10	70:30	25	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	35,00%
11	70:30	30	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	36,68%
12	70:30	35	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	34,30%
13	70:30	40	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	35,70%
14	70:30	45	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	39,40%
15	70:30	50	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	40,14%
16	**70:30**	**60**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**42,33%**
17	80:20	20	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	36%
18	80:20	25	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	38,46%
19	80:20	30	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	36,26%
20	**80:20**	**35**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**38,46%**
21	80:20	40	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	34%
22	80:20	45	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	31%
23	80:20	50	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	33%
24	80:20	60	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	32%

Setelah pengujian nilai K selesai didapat nilai akurasi terbaik dengan data set 60:40 nilai k=45 akurasi didapat sebesar 41,21%, data set 70:30 nilai k=60 didapatkan akurasi 42,33% dan data set 80:20 dengan nilai k=35 didapatkan akurasi 38,46%. Setelah pengujian nilai k selesai kemudian dilanjutkan dengan skenario pendekatan perilaku sosial dengan data set dan nilai k yang sudah didapatkan. Berikut tabel hasil akurasi skenario pendekatan perilaku sosial.

Tabel 10 Hasil Akurasi pada Skenario Pendekatan Perilaku Sosial

No	Data Set	K	Atribut											Akurasi
			FR	FG	MU	URL	M	RT	#	HB	TB	EM	KT	
1	60:40	45	Y	Y	Y	Y	Y							41,21%
2	**60:40**	**45**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**						**43,63%**
3	60:40	45	Y	Y	Y	Y	Y		Y					43,03%
4	60:40	45	Y	Y	Y	Y	Y			Y				35,75%
5	60:40	45	Y	Y	Y	Y	Y				Y			36,36%
6	60:40	45	Y	Y	Y	Y	Y					Y		35,15%
7	60:40	45	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	41,21%
8	70:30	60	Y	Y	Y	Y	Y							37,95%
9	70:30	60	Y	Y	Y	Y	Y	Y						38,68%
10	70:30	60	Y	Y	Y	Y	Y		Y					37,95%
11	70:30	60	Y	Y	Y	Y	Y			Y				40,14%
12	70:30	60	Y	Y	Y	Y	Y				Y			38,68%
13	70:30	60	Y	Y	Y	Y	Y					Y		40,14%
14	**70:30**	**60**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**41,60%**
15	80:20	35	Y	Y	Y	Y	Y							35,16%
16	80:20	35	Y	Y	Y	Y	Y	Y						37%
17	80:20	35	Y	Y	Y	Y	Y		Y					37%
18	80:20	35	Y	Y	Y	Y	Y			Y				34,06%
19	80:20	35	Y	Y	Y	Y	Y				Y			35,16%
20	80:20	35	Y	Y	Y	Y	Y					Y		31,86%

21	**80:20**	**35**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	**38%**

Dari hasil percobaan diatas hasil akurasi yang paling baik didapatkan 43,63% dengan data set 60:40 dan dengan fitur *follower, following, media url, url, mention, retweet* dengan nilai k sebesar 45. Berikut dibawah ini adalah hasil percobaan dengan pendekatan linguistik dengan menggunakan TF-IDF dan TF-*Chi Square.*

Tabel 11 Hasil Akurasi TF-IDF

TF-IDF			
No	**Data Set**	**K**	**Akurasi**
1	60:40	15	36,36%
2	60:40	20	38,18%
3	60:40	25	35,75%
4	60:40	30	34,54%
5	60:40	35	34,54%
6	**60:40**	**40**	**40%**
7	60:40	45	36,36%
8	60:40	50	36,96%
9	60:40	55	37,57%
10	60:40	60	36,96%
11	70:30	15	36,40%
12	70:30	20	37,22%
13	70:30	25	34,30%
14	70:30	30	35,06%
15	70:30	35	34,30%
16	70:30	40	34,30%
17	70:30	45	36,49%
18	**70:30**	**50**	**40,87%**
19	70:30	55	40,16%
20	70:30	60	37,92%
21	80:20	15	34,06%
22	80:20	20	34,06%
23	80:20	25	31,86%
24	80:20	30	31,86%
25	80:20	35	34,06%
26	80:20	40	32,96%
27	**80:20**	**45**	**34,06%**
28	80:20	50	32,96%
29	80:20	55	32,96%
30	80:20	60	31,86%

Tabel 12 Hasil Akurasi TF-*Chi Square*

TF-CHI SQUARE			
No	**Data Set**	**K**	**Akurasi**
1	60:40	15	39,39%
2	60:40	20	36,96%
3	60:40	25	38,18%
4	60:40	30	38,78%
5	60:40	35	37,57%
6	**60:40**	**40**	**43,03%**
7	60:40	45	41,21%
8	60:40	50	38,78%
9	60:40	55	41,21%
10	60:40	60	40,60%
11	**70:30**	**15**	**42,33%**
12	70:30	20	38,68%
13	70:30	25	35,57%
14	70:30	30	38,68%
15	70:30	35	38,68%
16	70:30	40	36,49%
17	70:30	45	37,95%
18	70:30	50	37,22%
19	70:30	55	37,22%
20	70:30	60	37,95%
21	80:20	15	30,76%
22	80:20	20	35,16%
23	80:20	25	34,06%
24	80:20	30	31,86%
25	80:20	35	35,16%
26	80:20	40	30,76%
27	80:20	45	35,16%
28	**80:20**	**50**	**36,26%**
29	80:20	55	34,06%
30	80:20	60	32,96%

Dari tabel 11 dan 12 merupakan hasil percobaan melalui pendekatan linguistik TF-IDF dan TF-*Chi Square*, hasil percobaan didapatkan hasil akurasi untuk setiap nilai k dan data set yang digunakan.

4.3. Hasil Uji

Dari seluruh percobaan dengan menggunakan beberapa fitur dan percobaan penggunaan rasio data set, didapat hasil yang berbeda-beda. Pada pendekatan perilaku sosial didapatkan nilai akurasi terbaik sebesar 43.63% pada fitur *follower, following*, media url, total url, *mention* dan *retweet* dengan perbandingan data latih dan data uji sebesar 60:40 dan dengan nilai k=45.

Pada pendekatan linguistik dengan menggunakan TF-IDF dengan rasio data set 60:40 dan nilai k 40 nilai akurasi sebesar 40% dan untuk data set 70:30 dengan nilai k 50 nilai akurasi sebesar 40,87% dan untuk data set 80:20 dengan nilai k 45 nilai akurasi yang didapat senilai 34,06%.

Pada pendekatan linguistik dengan menggunakan TF *Chi-Square* dengan rasio data set 60:40 dan nilai k 40 nilai akurasi 43,03% dan untuk data set 70:30 dengan nilai k 15 nilai akurasi yang didapat 42,33% dan untuk data set 80:20 dengan nilai k 50 mendapatkan akurasi senilai 36,26%.

5. Kesimpulan

Pada penelitian ini tentang prediksi kepribadian DISC pengguna tiwitter dengan menggunakan metode *K-Nearest Neighbors*(KNN) dengan menggunakan pembobotan TF-IDF dan TF-*Chi Square* yang bertujuan untuk memperoleh prediksi yang berakurasi tinggi yang diharapkan alternatif lain untuk menilai kepribadian seseorang. Dari hasil penelitian dapat diambil kesimpulan bahwa prediksi kepribadian DISC dengan metode KNN dan dengan dua fitur yaitu pendekatan perilaku sosial serta pendekatan linguistik didapatkan hasil akurasi untuk klasifikasi perilaku sosial sebesar 43,63% dengan k=45 dan untuk klasifikasi dengan linguistik mendapatkan akurasi sebesar 43,03% dengan k= 40 dan menggunakan pembobotan TF-*Chi Square*. Dari hasil yang didapat bisa disimpulkan pembobotan dengan TF-IDF dan TF-*Chi Square* tidak terlalu mempengaruhi akurasi terbukti dengan nilai akurasi pendekatan perilaku sosial lebih baik dari pada linguistik. Meskipun data yang di dapat berasal dari hasil psikotes mahasiswa Telkom tahun 2013, tetapi masih banyak didapatkan data yang berketimpangan cukup signifikan. Kelas *Steadiness* dan *Compilance* mempunyai anggota yang cukup banyak dari kelas lainnya yang menyebabkan akurasi pada model prediksi yang didapat tidak seimbang.

Beberapa alasan lain yang bisa mempengaruhi hasil akurasi seperti masih adanya tweet yang masih memiliki retweet dan masuk kedalam pengolahan *pre-processing* dan pengambilan kepribadian yang paling dominan juga bisa mengakibatkan nilai akurasi yang cukup rendah.

Saran untuk penelitian selanjutnya adalah untuk mencari algoritma yang bisa menangani masalah ketimpangan data agar hasil agar keputusan model yang dibuat tidak cenderung ke kelas yanng paling dominan. Lalu *tweet* yang di olah harus sudah bersih dari *retweet* supaya *tweet* yang olah adalah *tweet* pengguna itu sendiri, dan penentuan kelas dari data juga diperhatikan untuk membantu mendapatkan hasil yang lebih maksimal.

Daftar Pustaka

[1] J. Golbeck, C. Robles, and K. Turner, “Predicting personality with social media,” *Proc. 2011 Annu. Conf. Ext. Abstr. Hum. factors Comput. Syst. - CHI EA ’11*, p. 253, 2011.

[2] S. Asur and B. A. Huberman, “Predicting the Future with Social Media,” *2010 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, pp. 492–499, 2010.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter , a Social Network or a News Media?,” *Int. World Wide Web Conf. Comm.*, pp. 1–10, 2010.

[4] 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.

[5] S. Susanto and D. Suryani, “Pengantar Data Mining,” 2010.

[6] M. Certified, S. Engineer, and D. Mining, “D a t a M i n i n g,” pp. 1–3, 2003.

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

[8] A. D. Putri, “Klasifikasi Dokumen Teks Menggunakan Metode Support Vector Machine Dengan Pemilihan Fitur Chi-Square,” *skripsi. Inst. Pertan. Bogor, Bogor*, 2013.

[9] H. Leidiyana, “Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor,” *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[10] M. Fitri, “Perancangan Sistem Temu Balik Informasi Dengan Metode Pembobotan Kombinasi Tf-Idf Untuk Pencarian Dokumen Berbahasa Indonesia,” *J. Sist. dan Teknol. Inf.*, vol. 1, no. 1, pp. 1–6, 2013.

[11] D. I. Komputer, F. Matematika, and D. A. N. Ilmu, “Perbandingan metode seleksi fitur pada spam filter menggunakan klasifikasi multinomial naïve bayes julius gigih dimastyo,” 2014.

[12] U. K. Petra, “2. Kajian Pustaka,” *Peranc. Inter. Pus. Mitigasi di Jogja*, pp. 9–35, 2013.

[13] “Uji Chi-Square”, [Online]. Available: eprints.undip.ac.id/6796/1/CHI-KUADRAT.pdf. [Accessed: 11-Oct- 2018].

[14] H. Wu, “Reducing Over-Weighting in Supervised Term Weighting for Sentiment Analysis,” pp. 1322–1330, 2014.