Jurnal Teknik Informatika (JUTIF)
Vol. 3, No. 3, Juni 2022, hlm. 673-680

DOI: https://doi.org/10.20884/1.jutif.2022.3.3.254
p-ISSN: 2723-3863
e-ISSN: 2723-3871

COMPARISON NAÏVE BAYES CLASSIFIER, K-NEAREST NEIGHBOR AND SUPPORT VECTOR MACHINE IN THE CLASSIFICATION OF INDIVIDUAL ON TWITTER ACCOUNT

Aristin Chusnul Khotimah[*1], Ema Utami[2]

[1]Informatika Universitas Amikom Yogyakarta, Indonesia
[2]Magister Teknik Informatika Universitas Amikom Yogyakarta, Indonesia
Email: [1]aristin.khotimah@students.amikom.ac.id, [2]ema.u@amikom.ac.id

(Naskah masuk: 27 Maret 2022, Revisi : 30 Maret 2022, diterbitkan: 28 Juni 2022)

Abstract

In current's digital era, people can take advantage of the ease and effectiveness of interacting with each other. The most popular online activity in Indonesia is the use of sosial media. Twitter is a social media that allows people to build communication between users and get the latest information or news. Information obtained from twitter can be processed to get the characteristics of a person using the DISC method, DISC is a behavioral model that helps every human being why someone does. To classify the tweet into the DISC method using algorithms naïve bayes classifier, k-nearest neighbor and support vector machine with the TF-IDF weighting. The results is compare the accuracy of the naïve bayes classifier algorithm has an accuracy rate of 31.5%, k-nearest neighbor has an accuracy rate of 23.8%, while the support vector machine has an accuracy rate of 28.4%.

Keywords: *Personality DISC, Naïve Bayes Classifier, K-Nearest Neighbors, Support Vector Machine, Twitter*.

PERBANDINGAN ALGORITMA NAÏVE BAYES CLASSIFIER, K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DALAM KLASIFIKASI KARAKTER INDIVIDU PADA AKUN TWITTER

Abstrak

Pada era digital saat ini, masyarakat dapat memanfatkan kemudahan dan keefektifan dalam berinteraksi antara satu sama lain. Kegiatan online yang popular di Indonesia adalah penggunaan media sosial. Twitter adalah media sosial yang memungkinkan orang untuk membangun komunikasi antar pengguna dan dapatkan informasi terbaru atau berita. Informasi yang didapatkan dari twitter dapat diolah untuk mendapatkan karakteristik seseorang menggunakan metode DISC, DISC merupakan model perilaku yang membantu setiap manusia mengapa seseorang melakukan apa yang dia lakukan. Untuk mengklasifikasikan tweet pada akun twitter kedalam metode DISC menggunakan algoritma naïve bayes classifier, k-nearest neighbor dan support vector machine dengan pembobotan TF-IDF. Hasil dari membandingkan tingkat akurasi algoritma naïve bayes classifier memiliki tingkat akurasi sebesar 31.5%, algoritma k-nearest neighbor memiliki tingkat akurasi sebesar23.8%, dan algoritma support vector machine memiliki tingkat akurasi sebesar 28.4%.

Kata kunci: *Kepribadian DISC, Naïve Bayes Classifier, K-Nearest Neighbors, Support Vector Machine, Twitter.*

1. PENDAHULUAN

Pada era digital saat ini, masyarakat dapat memanfaatkan kemudahan dan keefektifan dalam berinteraksi antara satu sama lain. Kegiatan online yang popular di Indonesia adalah pengguna media sosial [1]. Media sosial merupakan aktivitas sosial yang menggunakan jaringan *online* berupa bahasa, gambar dan video (Purawinangun, 2020) Aktivitas sosial tersebut dianggap sangat memudahkan seseorang dalam berdiskusi, berbisnis, dan berkomentar secara bebas. (Isnain, Sihabuddin dan Suyanto, 2020) [2]. Perkembangan media sosial yang kini digunakan melengkapi dan memudahkan banyak pekerjaan, khususnya di dalam hal berbagi informasi dan menjalani komunikasi dengan banyak pihak. Komunikasi dan berbagai informasi tersebut didukung dengan munculnya media sosial seperti *twitter, facebook, tumblr, blog, instagram* dan lainnya yang kerap digunakan penggunanya

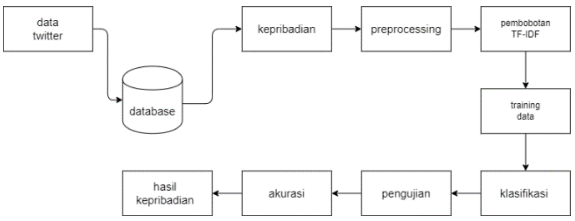
memperoleh informasi serta saling berbagi tanpa dibatasi ruang dan waktu [3].

Twitter adalah media sosial yang memungkinkan orang membangun komunikasi antar pengguna dan dapatkan informasi terbaru atau berita [4]. *Twitter* sering dijadikan objek penelitian dikarenakan sifatnya yang cepat karena dibatasinya jumlah karakter menjadi maksimal 280 karakter, sehingga lebih langsung menyampaikan maksud pesan. Twitter juga bermanfaat buat perusahan mencari informasi kepada calon rekrutmen karyawan. *Tweet* yang di posting seseorang bisa menggambarkan perasaan dan karakter dari orang tersebut. Karakter adalah kepribadian dari seseorang yang biasanys terbentuk darri internalisasi diri yang biasanya terbentuk dari internalisasi diri yang di yakinkan menjadi dasar dari cara pandang seseorang, cara berpikir, berperilaku, dan cara bertindak. Dalam ilmu psikolog, dimaknai sebagai hasil proses psikologis perkembangan seseorang yang dapat menjelaskan perbedaan antara satu individu dengan individu lainnya, serta menggambarkan perilaku alami manusia.

Melakukan tes kepribadian dapat dilakukan dengan menggunakan metode DISC, MBTI, *Strength Finder* dan *Big Five*. Namun, dalam penelitian ini, menggunakan pendekatan metode DISC. DISC ditemukan oleh seorang ahli psikoloh asal Amerika yang Bernama William Moulton Marston pada tahun 1928 dalam bukunya yang berjudul *Emotions of Normal People*. DISC adalah model perilaku yang membantu setiap manusia mengapa seseorang melakukan apa yang dia lakukan. DISC membagi karakter seseorang terdiri dari 4 tipe kepribadian yaitu, *Domainance (D), Influence (I), Steadiness (S), dan Compliance (C)* [6].

Namun melakukan hal ini, diperlukan sebuah metode lain yaitu *data mining*. *Data mining* menyangkut database, kecerdasan buatan, statistik. Untuk mengklasifikasikan *tweet* dari seseorang diperlukan algoritma yang terdapat dalam *data mining*. Pada penelitian ini membandingkan kinerja antara algoritma *naïve bayes classifier, k-nearest neighbor,* dan *support vector machine*.

2. METODE PENELITIAN



Gambar 1. Flowchart pada Sistem

2.1. Crawling

Crawling data merupakan tahap penelitian yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu *database*. Pengumpulan data dari penelitian ini yaitu data yang diunduh dari server *twitter* berupa user dan *tweet* beserta atribut-atributnya. Crawling data dibuat dengan memodifikasi *Application Programming Integration* (API) *twitter* dengan menggunakan bahasa *python* [8].

Application Programming Integration (API) *twitter* merupakan suatu program atau aplikasi yang disediakan oleh *twitter* untuk mempermudah *developer* lain dalam mengakses informasi yang ada di *website twitter*. Pendaftaran sebagai *developer* aplikasi *twitter* untuk menggunakan API *twitter* dapat dilakukan di lama *https://dev.twitter.com*. Setelah mendaftar *developer* akan mendapatkan *consumer key*, *consumer access*, *access token* dan *access token secret* yang akan digunakan sebagai syarat *otentifikasi* dari aplikasi yang akan kita bangun. Tujuan dari *otentifikasi* adalah untuk hak akses *developer* dalam mengunduh data yang ada di *twitter* [7].

2.2. Kepribadian DISC

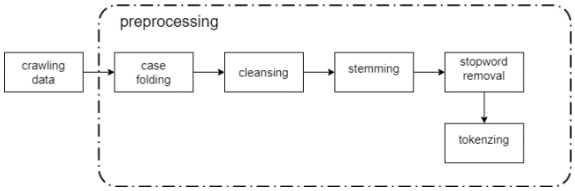
DISC pertama kali ditemukan oleh William Moulton Marston (1983-1947) seorang psikolog Amerika. DISC adalah sebuah alat ukur psikometri yang mengukur gaya kepribadian seseorang mengenai perilaku kerjanya [5]. DISC merupakan model perilaku yang membantu setiap manusia memahami. Hal ini kemudian akan menghasilkan empat kuadrat yang disertai dengan pola perilakunya yaitu [6]:

1. *Dominance (D)* Orang yang dominance tinggi bersifat asertif (tegas) dan langsung. Biasanya mereka sangat independent, ambisius, gagah serta menyukai tantangan dan persaingan. Dalam pemecahan masalahnya, melakukan pendekatan yang aktif dan cepat menyelesaikan masalah. Mereka dipandang orang lain sebagai orang yang berkemauan keras. Oleh karena itu mereka menginginkan segala sesuatu sesuai dengan kemauan mereka.
2. *Influence (I)* tipe influence ini senang berteman. Mereka suka menghibur orang lain dan bersifat sosial. Dalam penyelesaian masalah atau menghibur orang lain dan bersifat sosial. Dalam penyelesaian masalah atau mengerjakan sesuatu, mereka banyak mengandalkan keterampilan sosial. Orang bersifat interpersonal ini senang berpartisipasi dalam kelompok dan suka bekerja sama. Keterbukaan sikapnya membuat orang lain memandang dirinya sebagai pribadi yang gampang bergaul dan ramah. Biasanya pribadi ini memiliki banyak teman.
3. *Steadiness (S)* orang yang bertipe ini adalah orang yang pekerja keras hati, gigih, dan sabar. Mereka mendekati dan menjalani kehidupan dengan memanfaatkan standar yang terukur dan stabil. Pada umumnya mereka tidak begitu

suka kejutan dan tidak suka banyak menuntut dan bersifat akomodatif. Mereka sangat ramah dan memperlihatkan kesetiannya kepada disekitarnya. Orang yang bertipe seperti ini jujur dan sabar. Orang lain memandang mereka sebagai orang yang tenang, berhati-hati dan konsisten dalam cara mereka menjalani kehidupan.

4. *Compliance (C)* orang yang bertipe ini sangat tertarik pada presisi (ketelitian dan kecermatan) dan juga dengan akurasi (kecepatan). Mereka sangat focus terhadap fakta. Orang tipe teliti ini sangat menghargai peraturan. Dalam beraktivitas mereka menggunakan sistematis dan aturan agar semuanya terkelola dengan baik. Mengatasi konflik secara tidak langsung. Dihadapan orang lain, mereka dipandang pasif dan selalu mengalah.

2.3. Preprocessing



Gambar 2. Flowchar pada Preprocessing

2.3.1. Case Folding

Case folding dilakukan untuk mengubah huruf besar (*uppercase*) menjadi huruf kecil (*lowercase*) [10]. Hanya huruf a sampai dengan z yang diterima. Tujuan dari proses ini dilakukan untuk mempermudah dalam melakukan proses selanjutnya. Tidak semua dokumen teks konsisten dalampenggunaan huruf kapital [9].

2.3.2. Cleansing

Cleansing dilakukan untuk menghilangkan *tweet* dari kata yang kurang dibutuhkan *seperti URL, hashtag '#', dan username '@'*. Sedangkan *URL* ditandai dengan munculnya format *URL* seperti *http, https, atau www* [10]. Tujuan dari proses ini dilakukan untuk mengurangi *noise*.

2.3.3. Stemming

Stemming digunakan untuk mencari kata hubung, kata depan, atau kata ganti dengan menghilangkan imbuhan menjadi kata dasar. Pada proses ini teks berbahasa indonesia berbeda dengan teks berbahasa inggris. Pada teks berbahasa inggris, proses yang diperlukan hanya proses menghilangkan *sufiks*. Sedangkan teks berbahasa indonesia semua kata imbuhan baik itu *sufiks* dan *prefiks* juga dihilangkan [14].

2.3.4. Stopword Removal

Stopword removal digunakan penyaringan setiap kata, jika di dalam kata tersebut terdapat kata yang tidak memiliki makna, maka kata tersebut akan dihilangkan yang tidak ada hubungannya dalam krakter seseorang [9].

2.3.5. Tokenzing

Tokenzing digunakan untuk memperiksa kalimat *tweet* secara menyeluruh. Kemudian dilakukan pemenggalan kata berdasarkan karakter pemisahnya, sehingga kata yang bukan karakter pemisah akan digabungkan dengan karakter selanjutnya [16].

2.4. Pembobotan TF-IDF

Pembobotan fitur merupakan sebuah proses pemberian nilai pada setiap trem berdasarkan *relevansi* dan pengaruhnya terhadap hasil klasifikasi. Nilai tersebut nantinya dapat digunakan sebagai dasar untuk melakukan seleksi *trem* berdasarkan minimum bobot yang telah dihitung dari setiap *trem*. Pembobotan dilakukan dengan menggunakan metode TF-IDF. Algoritma TF-IDF pertama kali dicetuskan oleh Salton dan Buckley pada tahun 1988 dan digunakan untuk kepentingan *information retrieval*, yang kemudian turut dimanfaatkan sebagai salah satu algoritma dalam metode *feature weighting* dalam *text mining* [12]. TF-IDF memiliki persamaan sebagai berikut:

$$TF = \frac{Jumlah\ Kemunculan\ trem\ pada\ satu\ dokumen}{Jumlah\ seluruh\ trem\ dalam\ dokumen} \quad (1)$$

$$IDF = \frac{Jumlah\ seluruh\ dokumen}{Jumlah\ dokumen\ suatu\ trem\ muncul} \quad (2)$$

Semakin sering sebuah trem muncul, maka semakin besar pula bobot yang akan didapat artinya akan semakin penting pula *trem* tersebut.

2.5. Naive Bayes Classifier

Naïve bayes adalah salah satu algoritma yang popular digunakan untuk keperluan *data mining* karena kemudahan penggunaannya serta pemrosesan yang cepat, mudah diimplementasikan dengan strukturnya yang cukup sederhana dan tingkat efektifitas yang tinggi [13].

Naïve bayes merupakan algoritma yang dapat mengklasifikasi suatu variabel tertentu dengan menggunakan metode probabilitas dan statistic [15]. Bergantung pada model probolitasnya, *naïve bayes classifier* dapat dilatih untuk melakukan *supervised learning* dengan sangat efektif. Naïve bayes tidak membutuhkan jumlah data *training* yang banyak.

Naïve bayes menghitung peluang masuknya sampel karakter tertentu dalam kelas h (*posterior*) yaitu peluang munculnya kelas h dikali dengan kemunculan karakter sampel pada kelas c

(*likelihood*). Adapun bentuk umum dari *naïve bayes classifier* tertera pada persamaan:

$$P(H|X) = \frac{(P(X|H) \times P(H))}{P(X)} \quad (3)$$

Keterangan:
X = data dengan kelas yang belum
H = hipotesis data X merupakan suatu kelas spesifik
P(H|X) = probabilitas hipotesis H berdasar kondisi X (posterioriprobability)
P(H) = probabilitas hipotesis H (priorprobability)
P(X|H) = probabilitas X berdasarkan kondisi pada hipotesis H

2.6. K-Nearest Neighbor

K-Nearest Neighbor merupakan algoritma yang sering digunakan untuk klasifikasi teks dan data. Tujuan dari algoritma ini adalah mengklasifikasikan *obyek* berdasarkan atribut dan training sampel. Salah satu metode yang menerapkan algoritma *supervised*. Perbedaan antara *supervised* dan *unsupervised learning* adalah pada *supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada, dan sedangkan *unsupervised learning*, data belum memiliki pola apapun [5].

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

$$D(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \quad (4)$$

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

2.7. Support Vector Machine

Support Vector Machine adalah salah satu metode klasifikasi menggunakan machine learning (*supervised learning*). Konsep kerja *support vector machine* yaitu dengan mencari hyperplance atau garis pembatas paling optimal yang berfungsi untuk memisahkan dua kelas.

Untuk memperoleh garis hyperplance yang paling optimal dalam memisahkan data ke dua kelas tersebut, maka digunakan perhitungan *margin hyperplance* dan menemukan titik maksimal. Dalam memperoleh *hyperplane* pada *support vector machine*, dapat menggunakan persamaan:

$$(w \,.\, x_i) + b = 0 \quad (5)$$

Di dalam data x_i, yang termasuk pada kelas -1 dapat dirumuskan seperti persamaan:

$$(w \,.\, x_i) + b \leq 1, y_i = -1 \quad (6)$$

Sedangkan data data x_i, yang termasuk pada kelas +1 dapat dirumuskan seperti persamaan:

$$(w \,.\, x_i) + b \geq 1, y_i = 1 \quad (7)$$

Dalam proses klasifikasi dengan *support vector machine* biasanya ditemui kondisi dimana *kernel linear* bekerja tidak optimal yang mengakibatkan hasil klasifikasi terhadap data menjadi buruk. Hal tersebut dapat diatasi dengan menggunakan *kernel non-linear* dengan memanfaatkan *kernel trick*, akan dilakukan mapping data input ke *feature space* yang dimensinya lebih tinggi sehingga membuat data input yang dihasilkan akan terpisah secara *linear* dan membentuk *hyperplane* yang optimal [11]. Persamaan dari setiap karnel support vector machine dapat dilihat pada tabel:

Tabel 1. Persamaan Setiap Kernel

Nama Kernel	Definisi Fungsi
Linier	$\mathbf{K(x, y) = x \cdot y}$
Polynomial	$\mathbf{K(x, y) = (x \cdot y)^d}$
Gaussian (RBF)	$\mathbf{K(x, y) = exp\left(\frac{||x - y||^2}{2\sigma^2}\right)}$
Sigmoid	$\mathbf{K(x, y) = tan(\sigma(x, y) + c)}$

2.8. Confusion Matrix

Confusion matrix merupakan suatu instrumen yang digunakan untuk mengevaluasi performa dari model klasifikasi yang telah dihasilkan. Pada *confusion matrix*, hasil kelas prediksi akan dibangdingkan dengan kelas data yang sebenarnya. Hasil tersebut dibandingkan dengan kelas data yang sebenarnya. Hasil tersebut kemudian akan digunakan untuk menghitung nilai *accuracy*, *precision*, *recall* dan *f-score* [11]. Pengukuran evaluasi dengan *confusion matrix* dapat dilihat pada tabel:

Tabel 2. Confusion Matrix

Data Aktual	Data Prediksi		
	TRUE	FALSE	TOTAL
TRUE	TP	FN	P
FALSE	FP	TN	N
TOTAL	P'	N'	P+N

Keterangan:
TP = data positif yang terklasifikasi secara benar.
TN = data negatif yang terklasifikasi secara benar.
FP = data negatif yang terklasifikasi menjadi positif.
FN = data positif yang terklasifikasi menjadi negatif.

3. HASIL DAN PEMBAHASAN

Setalah data terkumpul melalui proses klasifikasi, 130 akun *twitter* akan divalidasi dan dievaluasi. Dari dataset akun *twitter* yang tervalidasi terdapat 130 akun. Data tersebut kemudian dikelompokkan secara manual berdasarkan label yang dilakukan oleh psikolog dan klasifikasi menggunakan algoritma *naïve bayes classifier, k-nearest neighbor, dan support vector machine.*

3.1. Algoritma Naïve Bayes Classifier

Tabel 3. Hasil Pengelompokan Psikolog dan Pengklasifikasi Data dengan Algoritma Naïve Bayes Classifier

Label	Akun berdasarkan Psikolog	Akun berdasarkan Naïve Bayes Classifier
Dominance (D)	51	87
Influence (I)	35	0
Steadiness (S)	13	37
Compliance (C)	31	6
Total Akun	130	130

Berdasarkan tabel terlihat bahwa menggunakan algoritma *naïve bayes classifier*, jumlah akun dengan kategori kepribadian *Dominance (D)* lebih banyak dikelompokkan, sedangkan kategori kepribadian lainnya lebih kecil, namun untuk kepribadian *Influence (I)* tidak memiliki jumlah pengkelompokan sama sekali.

Untuk mendapatkan hasil algoritma *naïve bayes classifier* yang lebih komprehensif pada data yang diperoleh dari psikolog, maka akan dilakukan pengujian data menggunakan metode *confusion matrix*.

Kinerja diperoleh dengan memberikan nilai pada confusion untuk menghitung nilai *akurasi*, *presisi*, dan *recall* dari hasil pengujian. Berikut ini *confusion matrix* dijelaskan:

Tabel 4. Confusion Matrix Algoritma Naïve Bayes Classifier

		Aktual			
		D	I	S	C
Hasil dari Psikolog	D	33	0	15	3
	I	24	0	10	1
	S	5	0	7	1
	C	25	0	5	1

Kinerja dengan menggunakan *confusion matrix* memiliki empat kemungkinan keluaran sebagai respresentasi dari hasil proses klasifikasi, yaitu *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). True Positive (TP)* adalah jumlah sebenarnya dari data positif. *True Negative (TN)* adalah jumlah sebenarnya dari data negative. *False Positive (FP)* adalah jumlah semua kolom untuk setiap kelas, kecuali nilai *True Positive (TP). False Negative (FN)* adalah jumlah semua baris untuk setiap kelas, kecuali nilai *True Positive (TP).* Dengan menggunakan metode evaluasi ini, tingkat akurasi keseluruhan kinerja *naïve bayes classifier* dapat dihitung sebagai berikut:

$$akurasi = \frac{33+0+7+1}{130} \times 100\%$$

$$akurasi = \frac{41}{130} \times 100\%$$

$$akurasi = 0.315 \times 100\%$$

$$akurasi = 31.5\%$$

Dari hasil perhitungan nilai akurasi ternyata menggunakan algoritma pengklasifikasi *naïve bayes classifier* tidak begitu baik yaitu 31.5% dengan membandingkan hasil klasifikasi sistem dengan hasil psikolog.

Selain perhitungan *akurasi*, *nilai presisi* dan *recall* untuk setiap label juga dapat dihitung. *Presisi* atau nilai prediksi positif adalah tingkat keakuratan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem. Sedangkan *recall* adalah tingkat keberhasilan sistem dalam menemukan Kembali informasi. Untuk hasil perhitungan *presisi* dan *recall* dari masing-masing label dijelaskan:

Tabel 5. Presisi dan Recall Algoritma Naïve Bayes Classifier

Label	Presisi	Recall
D	37.9%	64.7%
I	∞	0%
S	18.9%	53.8%
C	16.6%	3.22%

3.2. Algoritma K-Nearest Neighbor

Tabel 6. Hasil Pengelompokan Psikolog dan Pengklasifikasi Data dengan Algoritma K-Nearest Neighbor

Label	Akun berdasarkan Psikolog	Akun berdasarkan K-Nearest Neighbor
Dominance (D)	51	13
Influence (I)	35	73
Steadiness (S)	13	39
Compliance (C)	31	5
Total Akun	130	130

Berdasarkan tabel terlihat bahwa menggunakan algoritma *k-nearest neighbor*, jumlah akun dengan kategori kepribadian *Influence (I)* lebih banyak dikelompokkan, sedangkan kategori kepribadian lainnya lebih kecil.

Untuk mendapatkan hasil algoritma *k-nearest neighbor* yang lebih komprehensif pada data yang diperoleh dari psikolog, maka akan dilakukan pengujian data menggunakan metode *confusion matrix*.

Kinerja diperoleh dengan memberikan nilai pada *confusion* untuk menghitung nilai *akurasi*, *presisi*, dan *recall* dari hasil pengujian. Berikut ini *confusion matrix* dijelaskan:

Tabel 7. Confusion Matrix Algoritma K-Nearest Neighbor

		Aktual			
		D	I	S	C
Hasil dari Psikolog	D	4	30	15	2
	I	3	21	10	1
	S	2	5	5	1
	C	4	17	9	1

Kinerja dengan menggunakan *confusion matrix* memiliki empat kemungkinan keluaran sebagai respresentasi dari hasil proses klasifikasi, yaitu *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). True Positive (TP)* adalah jumlah sebenarnya dari data positif. *True Negative (TN)* adalah jumlah sebenarnya dari data negative. *False Positive (FP)* adalah jumlah semua

kolom untuk setiap kelas, kecuali nilai *True Positive (TP). False Negative (FN)* adalah jumlah semua baris untuk setiap kelas, kecuali nilai *True Positive (TP)*. Dengan menggunakan metode evaluasi ini, tingkat akurasi keseluruhan kinerja k-nearest neighbor dapat dihitung sebagai berikut:

$$akurasi = \frac{4+21+5+1}{130} \times 100\%$$

$$akurasi = \frac{31}{130} \times 100\%$$

$$akurasi = 0.238 \times 100\%$$

$$akurasi = 23.8\%$$

Dari hasil perhitungan nilai akurasi ternyata menggunakan algoritma pengklasifikasi *k-nearest neighbor* tidak begitu baik yaitu 23.8% dengan membandingkan hasil klasifikasi sistem dengan hasil psikolog.

Selain perhitungan *akurasi*, nilai *presisi* dan *recall* untuk setiap label juga dapat dihitung. *Presisi* atau nilai prediksi positif adalah tingkat keakuratan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem. Sedangkan *recall* adalah tingkat keberhasilan sistem dalam menemukan Kembali informasi. Untuk hasil perhitungan *presisi* dan *recall* dari masing-masing label dijelaskan:

Tabel 8. Presisi dan Recall Algoritma K-Nearest Neighbor

Label	Presisi	Recall
D	23%	7.8%
I	28.7%	60%
S	12.8%	38.4%
C	20%	3.2%

3.3. Algoritma Support Vector Machine

Tabel 9. Hasil Pengelompokan Psikolog dan Pengklasifikasi Data dengan Algoritma Support Vector Machine

Label	Akun berdasarkan Psikolog	Akun berdasarkan Support Vector Machine
Dominance (D)	51	1
Influence (I)	35	96
Steadiness (S)	13	31
Compliance (C)	31	2
Total Akun	130	130

Berdasarkan tabel terlihat bahwa menggunakan algoritma *support vector machine*, jumlah akun dengan kategori kepribadian *Influence (I)* lebih banyak dikelompokkan, sedangkan kategori kepribadian lainnya lebih kecil.

Untuk mendapatkan hasil algoritma *support vector machine* yang lebih komprehensif pada data yang diperoleh dari psikolog, maka akan dilakukan pengujian data menggunakan metode *confusion matrix*.

Kinerja diperoleh dengan memberikan nilai pada *confusion* untuk menghitung nilai *akurasi*, *presisi*, dan *recall* dari hasil pengujian. Berikut ini *confusion matrix* dijelaskan:

Tabel 10. Confusion Matrix Algoritma Support Vector Machine

		Aktual			
		D	I	S	C
Hasil dari Psikolog	D	1	35	14	1
	I	0	29	6	0
	S	0	7	6	0
	C	0	25	5	1

Kinerja dengan menggunakan confusion matrix memiliki empat kemungkinan keluaran sebagai respresentasi dari hasil proses klasifikasi, yaitu *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). True Positive (TP)* adalah jumlah sebenarnya dari data positif. *True Negative (TN)* adalah jumlah sebenarnya dari data negative. *False Positive (FP)* adalah jumlah semua kolom untuk setiap kelas, kecuali nilai *True Positive (TP). False Negative (FN)* adalah jumlah semua baris untuk setiap kelas, kecuali nilai *True Positive (TP)*. Dengan menggunakan metode evaluasi ini, tingkat akurasi keseluruhan kinerja *support vector machine* dapat dihitung sebagai berikut:

$$akurasi = \frac{1+29+6+1}{130} \times 100\%$$

$$akurasi = \frac{37}{130} \times 100\%$$

$$akurasi = 0.284 \times 100\%$$

$$akurasi = 28.4\%$$

Dari hasil perhitungan nilai akurasi ternyata menggunakan algoritma pengklasifikasi *support vector machine* tidak begitu baik yaitu 28.4% dengan membandingkan hasil klasifikasi sistem dengan hasil psikolog.

Selain perhitungan *akurasi*, nilai *presisi* dan *recall* untuk setiap label juga dapat dihitung. *Presisi* atau nilai prediksi positif adalah tingkat keakuratan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem. Sedangkan *recall* adalah tingkat keberhasilan sistem dalam menemukan Kembali informasi. Untuk hasil perhitungan *presisi* dan *recall* dari masing-masing label dijelaskan:

Tabel 11. Presisi dan Recall Algoritma Support Vector Machine

Label	Presisi	Recall
D	100%	1.9%
I	30.2%	82.8%
S	19.3%	46.1%
C	50%	3.2%

4. KESIMPULAN

Penelitian ini menggunakan data akun *twitter* berjumlah 130 akun dengan masing-masing akun berjumlah 200 *tweet*. Akun *twitter* dan masing-masing *tweet* pada akun tersebut diberi pelabelan sebelumnya secara manual oleh seorang psikolog yang mampu menganalisis karakter seseorang. Berdasarkan penelitian diatas terdapat beberapa kesimpulan, Adapun kesimpulanya sebagai berikut yaitu Data *tweet* dari pengguna akun *twitter* didapatkan dengan memanfaatkan API yang

tersedia. Algoritma *naïve bayes classifier*, *k-nearest neighbor*, dan *support vector machine* telah berhasil mengklasifikasikan akun *twitter* kedalam karakter DISC (*Dominance, Influence, Steadiness, Conscitieous*) meskipun dengan tingkat *akurasi* yang cukup rendah.

Hasil perhitungan *evaluasi* yang telah dilakukan dengan *confusion matrix* bahwa tingkat *akurasi* yang didapat dari membandingkan hasil dari psikolog secara manual dan hasil klasifikasi dari sistem untuk algoritma *naïve bayes classifier* adalah 31.5%, algoritma *k-nearest neighbor* adalah 23.8%, sedangkan algoritma *support vector machine* adalah 28.5%. Dapat disimpulkan bahwa algoritma *naïve bayes classifier* lebih unggul dibandingkan algoritma *k-nearest neighbor*, dan *support vector machine*, namun algoritma *support vector machine* lebih unggul perbedaan sebesar 4.7% dengan algoritma *k-nearest neighbor.*

DAFTAR PUSTAKA

[1] Agustina, Dyah Auliya, Sri Subanti, and Etik Zukhronah. "Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine." *Indonesian Journal of Applied Statistics* 3.2 (2021): 109-122.

[2] Alita, Debby, Yusra Fernando, and Heni Sulistiani. "Implementasi Algoritma Multiclass SVM pada Opini Publik Berbahasa Indonesia di Twitter." *Jurnal Tekno Kompak* 14.2 (2020): 86-91.

[3] Srisadono, Wahyu. "Komunikasi Publik Calon Gubernur Provinsi Jawa Barat 2018 dalam Membangun Personal Branding Menggunakan Twitter." *Jurnal Pustaka Komunikasi* 1.2 (2018): 213-227.

[4] Hartanto, Anggit Dwi, et al. "Job seeker profile classification of twitter data using the naïve bayes classifier algorithm based on the DISC method." *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. IEEE, 2019.

[5] Munggaran, Nur Ihsan Putra, and Erwin Budi Setiawan. "Prediksi Kepribadian Disc Dengan K-nearest Neighbors Algorithm (knn) Menggunakan Pembobotan Tf-idf Dan Tf-chi Square." *eProceedings of Engineering* 6.2 (2019).

[6] Tambunan, Maulina Gustiani, and Erwin Budi Setiawan. "Prediksi Kepribadian Disc Pada Twitter Menggunakan Metode Decision Tree C4. 5 Dengan Pembobotan Tf-idf Dan Tf-rf." *eProceedings of Engineering* 7.1 (2020).

[7] Sembodo, J. Eka, E. Budi Setiawan, and Z. Abdurahman Baizal. "Data Crawling Otomatis pada Twitter." *Indonesian Symposium on Computing (Indo-SC)*. 2016.

[8] Ramadhan, Dery Anjas, and Erwin Budi Setiawan. "Analisis Sentimen Program Acara di SCTV pada Twitter Menggunakan Metode Naive Bayes dan Support Vector Machine." *eProceedings of Engineering* 6.2 (2019).

[9] Nugroho, Agung. "Analisis Sentimen Pada Media Sosial Twitter Menggunakan Naive Bayes Classifier Dengan Ekstrasi Fitur N-Gram." *J-SAKTI (Jurnal Sains Komputer dan Informatika)* 2.2 (2018): 200-209.

[10] Kurniawan, Imam, and Ajib Susanto. "Implementasi Metode K-Means dan NaÃ¯ve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019." *Jurnal Eksplora Informatika* 9.1 (2019): 1-10.

[11] Husada, Hendry Cipta, and Adi Suryaputra Paramita. "Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM)." *Teknika* 10.1 (2021): 18-26.

[12] Prameswari, Kartika, and Erwin Budi Setiawan. "Analisis Kepribadian Melalui Twitter Menggunakan Metode Logistic Regression dengan Pembobotan TF-IDF dan AHP." *eProceedings of Engineering* 6.2 (2019).

[13] Hadna, N. Muchammad Shiddieqy, P. Insap Santosa, and Wing Wahyu Winarno. "Studi literatur tentang perbandingan metode untuk proses analisis sentimen di Twitter." *Semin. Nas. Teknol. Inf. dan Komun* 2016 (2016): 57-64.

[14] H. Aria, W. Titin and I. Henny, "PENGEMBANGAN STEMMING UNTUK ARTIKEL BERBAHASA INDONESIA.", https://repository.usm.ac.id/files/research/G071/20200410024415-PENGEMBANGAN-STEMMING-UNTUK-ARTIKEL-BERBAHASA-INDONESIA.pdf, diakses pada 25 Februari 2022

[15] Edwin, Edwin. "APLIKASI ASSESMENT KEBIJAKAN PEMERINTAH TERKAIT OPERASI OJEK ONLINE DI MASA PANDEMI (COVID-19) MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER." *Jurnal Algoritma, Logika dan Komputasi* 3.2 (2021).

[16] Kurniawan, Robi, and Aulia Apriliani. "Analisis sentimen masyarakat terhadap virus corona berdasarkan opini dari Twitter berbasis web scraper." *Jurnal INSTEK (Informatika Sains dan Teknologi)* 5.1

(2020): 67-75.