TESIS

ANALISIS PROFIL PENGGUNA MEDIA SOSIAL BERBAHASA INDONESIA BERDASARKAN FRAMEWORK DISC DAN OCEAN MENGGUNAKAN SUPPORT VECTOR MACHINE

Disusun oleh:

Nama	**: Muhammad Ryandy Ghonim Asgar**
NIM	**: 21.51.2108**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2024

TESIS

ANALISIS PROFIL PENGGUNA MEDIA SOSIAL BERBAHASA INDONESIA BERDASARKAN FRAMEWORK DISC DAN OCEAN MENGGUNAKAN SUPPORT VECTOR MACHINE

ANALYSIS OF USER PROFILES OF INDONESIAN LANGUAGE SOCIAL MEDIA BASED ON DISC AND OCEAN FRAMEWORKS USING SUPPORT VECTOR MACHINE

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Muhammad Ryandy Ghonim Asgar**
NIM : **21.51.2108**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2024

HALAMAN PENGESAHAN

ANALISIS PROFIL PENGGUNA MEDIA SOSIAL BERBAHASA INDONESIA BERDASARKAN FRAMEWORK DISC DAN OCEAN MENGGUNAKAN SUPPORT VECTOR MACHINE

ANALYSIS OF USER PROFILES OF INDONESIAN LANGUAGE SOCIAL MEDIA BASED ON DISC AND OCEAN FRAMEWORKS USING SUPPORT VECTOR MACHINE

Dipersiapkan dan Disusun oleh

Muhammad Ryandy Ghonim Asgar

21.51.2108

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Senin, 5 Agustus 2024

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 5 Agustus 2024
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

ANALISIS PROFIL PENGGUNA MEDIA SOSIAL BERBAHASA INDONESIA BERDASARKAN FRAMEWORK DISC DAN OCEAN MENGGUNAKAN SUPPORT VECTOR MACHINE

ANALYSIS OF USER PROFILES OF INDONESIAN LANGUAGE SOCIAL MEDIA BASED ON DISC AND OCEAN FRAMEWORKS USING SUPPORT VECTOR MACHINE

Dipersiapkan dan Disusun oleh

Muhammad Ryandy Ghonim Asgar

21.51.2108

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Senin, 5 Agustus 2024

Pembimbing Utama	**Anggota Tim Penguji**
Prof. Dr. Ema Utami, S.Si., M.Kom	**Dr. Andi Sunyoto, M.Kom**
NIK. 190302037	**NIK. 190302052**
	Alva Hendi M, S.T., M.Eng., P.h.D.
Pembimbing Pendamping	**NIK. 190302493**
Ainul Yaqin, M.Kom	**Prof. Dr. Ema Utami, S.Si., M.Kom**
NIK. 190302255	**NIK. 190302037**

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 5 Agustus 2024
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Muhammad Rynady Ghonim Asgar
NIM : 21.51.2108
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Analisis Profil Pengguna Media Sosial Berbahasa Indonesia Berdasarkan Framework DISC dan OCEAN Menggunakan Support Vector Machine

Dosen Pembimbing Utama : Prof. Dr. Ema Utami, S.Si., M.Kom
Dosen Pembimbing Pendamping : Ainul Yaqin, M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 5 Agustus 2024
Yang Menyatakan,

Muhammad Ryandy Ghonim Asgar

HALAMAN PERSEMBAHAN

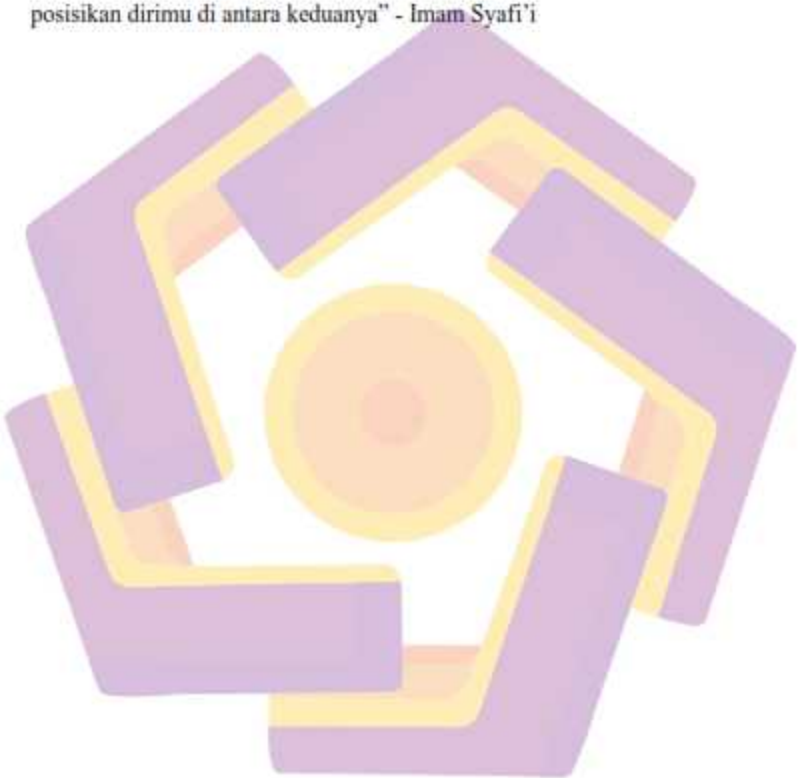
Pertama saya mengucapkan puja dan puji Syukur kepada Allah SWT atas berkat rahmat dan hidayah-Nya, sehingga penulis diberi kesempatan untuk menyelesaikan laporan tesis ini. Dengan begitu penulis mempersembahkan laporan tesis ini kepada:

1. Orang tua dan kakak saya, yang senantiasa memberikan dukungan, bantuan, semangat, motivasi, dan doa, semoga selalu berada dalam lindungan Allah SWT.
2. Ibu Prof. Dr. Ema Utami, S.Si., M.Kom., dan Pak Ainul Yaqin, S.Kom., M.Kom., yang telah memberikan bantuan serta bimbingan selama pelaksanaan penelitian, semoga mendapatkan keberkahan dan dilancarkan segala uruasannya.
3. Keluarga besar, teman, dan rekan kerja yang selalu memberikan dukungan dan doa selama ini.

Terimakasih atas semua semangat serta dukungan dari berbagai pihak. Semoga tesis ini dapat memberikan manfaat serta berguna dimasa yang akan datang.

HALAMAN MOTTO

"Terlalu keras dan menutup diri terhadap orang lain akan mendatangkan musuh, dan terlalu terbuka juga akan mendatangkan kawan yang tidak baik. Maka posisikan dirimu di antara keduanya" - Imam Syafi'i



KATA PENGANTAR

Puji syukur dipanjatkan kehadiran Allah ta'ala yang telah melimpahkan segala kebutuhan yang diperlukan selama penyusunan tesis ini, sehingga penulis dapat menyelesaikan penyusunan tesis dengan judul " Analisis Profil Pengguna Media Sosial Berbahasa Indonesia Berdasarkan Framework DISC dan OCEAN Menggunakan Support Vector Machine". Bimbingan dan bantuan berharga dari berbagai pihak tidak terlepas dari penyusunan tesis ini. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. M. Suyanto, MM., selaku Rektor Universitas Amikom Yogyakarta.
2. Prof. Dr. Ema Utami, S.Si., M.Kom., sebagai pembimbing utama yang telah mendampingi dalam penyusunan tesis ini.
3. Pak Ainul Yaqin, S.Kom., M.Kom., sebagai pendamping yang tak lelah membimbing penulis dalam pengerjaan tesis.
4. Bapak/Ibu dosen Magister Teknik Informatika Universitas Amikom Yogyakarta yang telah memberikan ilmunya kepada penulis selama menempuh studi Magister Teknik Informatika
5. Orang tua dan kakak yang memberikan dukungan serta doa.
6. Teman – teman MTI-2021 kelas A Universitas Amikom Yogyakarta yang telah memberikan dukungan dan doa.
7. Terakhir, kepada semua pihak yang telah membantu yang tidak dapat penulis sampaikan satu-persatu.

Penulis berharap penelitian ini dapat membantu bagi pihak yang membutuhkan. Kekurangan dan ketidaksempurnaan masih dapat ditemukan agar dapat menjadi patokan kearah yang lebih baik lagi dimasa yang akan datang. Selebihnya permohonan maaf apabila terdapat salah kata dan salah dalam penyusunan, Terima kasih

Yogyakarta, 21 Mei 2024

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Pemanfaatan data di sosial media Twitter untuk kepentingan penelitian telah banyak dilakukankarena adanya keterbukaan data yang disediakan dan berisi konten personal dan sosial dari pengguna yang dapat digunakan untuk melihat kebutuhan dan komentar atau emosi yang sedang dirasakan pengguna secara real-time. Salah satunya adalah dapat digunakan untuk mengenali ciri-ciri kepribadian seseorang berdasarkan tweet yang telah mereka unggah di Twitter. Kepribadian seseorang dapat diklasifikasikan kedalam kedua framework yaitu DISC atau OCEAN. Penelitian ini akan mengadopsi kedua framework tersebut untuk selanjutnya memanfaatkan metode Support Vector Machine untuk proses klasifikasi

Penelitian ini memiliki empat skenario untuk mendapatkan hasil yang optimal. Skenario pertama, penelitian dilakukan dengan memilih tahapan preprocessing yang tepat dengan memilih untuk menggunakan stopword removal atau tidak. Skenario kedua dilakukan dengan membagi dataset menjadi 4 bagian, yaitu 90:10, 80:20, 70:30, dan 60:40. Skenario ketiga dilakukan dengan memilih feature extraction yang meiliputi TF-IDF, WF-IDF dan Ngram. Terakhir dilakukan pemilihan kernel pada SVM, yaitu linear, rbf, sigmoid dan polinomial.

Hasil akhir penelitian ini menunjukkan bahwa SVM memiliki performa yang cukup baik dalam melakukan klasifikasi kepribadian menggunakan framework DISC dan OCEAN. Pada frameowrk DISC diperoleh akurasi 56%, presisi 63%, recall 36%, dan F1Score 34% dengan preprocessing data tanpa stopword removal, pembagian dataset dengan rasio 70:30, feature extraction n-gram, dan kernel SVM linear. Sedangkan pada dataset OCEAN diperoleh akurasi 67%, presisi 73%, recall 53%, dan F1Score 56% dengan preprocessing data tanpa stopword removal, pembagian dataset dengan rasio 60:40, feature extraction n-gram, dan kernel SVM linear

Kata kunci: Analisis Kepribadian, DISC, OCEAN, Support Vector Machine, Twitter

ABSTRACT

The use of data on Twitter social media for research purposes has been widely carried out because of the openness of the data provided and containing personal and social content from users which can be used to see needs and comments or emotions that users are feeling in real-time. One of them is that it can be used to recognize someone's personality traits based on the tweets they have uploaded on Twitter. A person's personality can be classified into two frameworks, namely DISC or OCEAN. This research will adopt these two frameworks to further utilize the Support Vector Machine method for the classification process

This research has four scenarios to obtain optimal results. In the first scenario, research is carried out by choosing the right preprocessing stage by choosing whether to use stopword removal or not. The second scenario is carried out by dividing the dataset into 4 parts, namely 90:10, 80:20, 70:30, and 60:40. The third scenario is carried out by selecting feature extraction which includes TF-IDF, WF-IDF and Ngram. Finally, the kernel selection is carried out on the SVM, namely linear, rbf, sigmoid and polynomial.

The final results of this research show that SVM has quite good performance in classifying personality using the DISC and OCEAN frameworks. In the DISC framework, accuracy of 56%, precision of 63%, recall of 36%, and F1Score of 34% were obtained with data preprocessing without stopword removal, dataset division with a ratio of 70:30, n-gram feature extraction, and linear SVM kernel. Meanwhile, on the OCEAN dataset, 67% accuracy, 73% precision, 53% recall, and 56% F1Score were obtained with data preprocessing without stopword removal, dataset division with a 60:40 ratio, n-gram feature extraction, and linear SVM kernel.

Keyword: Analys Profil, DISC, OCEAN, Support Vector Machine, Twitter

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Di era modern saat ini, perkembangan teknologi semakin pesat dan mengalami banyak kemajuan yang tentunya semakin memudahkan aktivitas manusia. Pengguna aktif internet di Indonesia pada Tahun 2022 mencapai 204 juta pengguna, meningkat sekitar 15.5% dari tahun 2021 yang hanya mencapai 202 juta pengguna, dan sekitar 170 juta pengguna menggunakan internet untuk membuka sosial media (Hootsuite, 2022). Ini menandakan bahwa lebih dari 50% penduduk Indonesia merupakan pengguna aktif internet dan aktif juga dalam sosial media. Sehingga tidak heran jika setiap perusahaan berlomba-lomba menyediakan sosial media sebagai wadah dari setiap individu untuk berhubungan dengan kerabat, bahkan orang yang tidak dikenal di kehidupan nyata. Namun tidak jarang juga sosial media digunakan sebagai penyaluran emosi dari setiap penggunanya. Tidak hanya itu, sosial media juga menjadi sumber informasi tercepat dan juga digunakan sebagai wadah untuk melakukan promosi tentang banyak hal (H. Krasnova et al, 2012). Sehingga dapat dikatakan bahwa sosial media merupakan *platform* paling penting di era modern saat ini.

Salah satu sosial media yang banyak digunakan di Indonesia adalah Twitter. Sama seperti sosial media pada umumnya, twitter memungkinkan penggunanya untuk menjalin komunikasi antar pengguna lainnya dan mendapatkan informasi, berita terbaru, hingga informasi terkait lowongan pekerjaan (Hartanto et al, 2019).

Berdasarkan hasil riset dari statista, pengguna Twitter di Indonesia mencapai 16.32 juta pada tahun 2021. Pemanfaatan data di sosial media Twitter untuk kepentingan penelitian telah banyak dilakukan. Hal ini dikarenakan keterbukaan data yang disediakan dan berisi konten personal dan sosial dari pengguna yang dapat digunakan untuk melihat kebutuhan dan komentar atau emosi yang sedang dirasakan pengguna secara real-time. Penelitian ini dirancang untuk melihat ciri-ciri kepribadian seseorang berdasarkan tweet yang telah mereka unggah di Twitter.

Kepribadian merupakan kualitas psikologis yang mempengaruhi pola perilaku karakteristik individu secara stabil dan bertahan lama, yang membedakan setiap individu dengan yang lainnya (Langford et al, 2020). Kepribadian setiap orang membuat seseorang unik dan berbeda dengan individu lain. Kepribadian seseorang terdiri dari karakteristik, pola pikir, perasaan dan perilaku, dimana sikap atau perilaku seseorang dapat dipengaruhi berdasarkan emosinya.

Pada tahun 2017 Nadeem dkk (Ahmad et al, 2017) melakukan penelitian dengan mengadopsi metode *text mining* dan analisis sentiment yang bertujuan untuk memprediksi karakteristik kepribadian pengguna Twitter dengan menggunakan *framework* DISC (*Dominance, Influence, Steadiness, Compliance*). Penelitian ini menunjukkan bahwa *framework* DISC mampu berjalan dengan baik dalam *text mining* untuk memprediksi kepribadian seseorang. Hasil akurasi terbaik yang didapatkan pada penelitian ini sebesar 35,6%.

Ortigosa dkk (Ortigosa et al, 2014) melakukan penelitian untuk menganalisis kepribadian seseorang. Dengan menggunakan sosial media Facebook sebagai dataset, dengan jumlah dataset yang digunakan sebanyak 20.000 users.

Pada penelitian ini menggunakan Teknik *machine learning* khususnya *Decision Tree* yang diaplikasikan untuk mengidentifikasi pola interaksi yang yang berkolerasi dengan ciri-ciri kepribadian user. Model klasifikasi dibangun dengan menggunakan 2 eksperimen, yaitu 3 model (low, medium, high) dan 5 model kelas (very high, high, medium, low, very low). Hasil terbaik pada penelitian ini mencapai 70% pada saat menggunakan 3 model kelas, sedangkan 62% untuk 5 model kelas. Metode ini dapat bermanfaat diberbagai bidang seperti e-commerce, e-learning, hingga assistive technologies. Selanjutnya Souri dkk (Souri et al, 2018) kembali melakukan penelitian untuk menganalisis kepribadian seseorang dengan menggunakan dataset yang sama seperti penelitian (Ortigosa et al, 2014). Pada penelitian ini Souri dkk menggunakan *framework five-factors* OCEAN (*Openness to experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism*) dengan menggunakan 2 kelas yaitu low dan high. Hasil dari penelitian ini berhasil meningkatkan akurasi decision tree menjadi 87% yang sebelumnya hanya mencapai 70% pada penelitian sebelumnya.

Penelitian untuk menganalisis kepribadian seseorang dilakukan dengan menggunakan Bahasa Indonesia. Ema Utami dkk (Utami et al, 2022) melakukan penelitian ini dengan tujuan untuk menganalisis apakah pengolahan tweet dari sosial media Twitter dapat menjadi alternatif HR (*Human Resource*) dalam menggali kepribadian calon karyawan. Dengan menggunakan *framework* DISC dan 4 skenario (*not-stemmed- not-weighted, stemmed-not weighted, not- stemmed-weighted, and stemmed- weighted*), penelitian ini berhasil mendapatkan nilai akurasi tertinggi pada skenario *not-stemmed- not-weighted* dengan akurasi 37,41%.

Hasil ini menunjukkan bahwa pendekatan ini memerlukan peningkatan dan pengembangan lebih lanjut untuk mencapai akurasi yang lebih tinggi dan efisien. Kemudian Hartanto dkk (Hartanto et al, 2019) kembali melakukan penelitian untuk memberikan alternatif kepada HR dalam melihat kepribadian calon karyawan melalui hasil tweet dari Twitter mereka. Penelitian ini menggunakan framework DISC dengan klasifikasi pembobotan W-IDF dan menambahkan algoritma Naive Bayes. Hasil penelitian menunjukkan bahwa meskipun Twitter dapat digunakan untuk mengidentifikasi kepribadian seseorang melalui postingan tweet mereka, akurasi menggunakan algoritma Naïve Bayes dan W-IDF masih tergolong cukup rendah, hanya 36,67%. Hal ini menunjukkan perlunya pengembangan lebih lanjut dalam hal metodologi, seperti peningkatan jumlah data yang telah dilabeli oleh pakar hingga akun Twitter yang telah divalidasi oleh pakar psikolog.

Penelitian tentang penggunaan *framework* DISC dan OCEAN sudah banyak dilakukan oleh peneliti lain, namun masih sangat jarang ditemukan penggunaan dataset dalam Bahasa Indonesia. Oleh karena itu, penelitian ini melakukan perbandingan *framework* DISC dan OCEAN pada postingan sosial media Twitter dalam Bahasa Indonesia untuk mengetahui kepribadian user berdasarkan tweet yang telah mereka unggah. Peneliti meyakini bahwa pentingnya melakukan perbandingan antara *framework* DISC dan OCEAN dalam analisis kepribadian seseorang, untuk mengetahui framework yang paling cocok digunakan dalam Bahasa Indonesia dan untuk mendukung penelitian lain kedepannya.

Untuk menghasilkan hasil yang akurat dalam proses klasifikasi kepribadian di sosial media seperti Twitter, diperlukan prosedur dan metode yang tepat untuk

melakukan kalsifikasi. Pada penelitian ini, metode yang akan digunakan adalah *Support Vector Machine* (SVM). Metode SVM dipilih karena memiliki performa yang baik dalam klasifikasi data dengan margin yang jelas antara kelas-kelas yang berbeda. Pada beberapa penelitian pada klasifikasi teks, SVM juga cenderung memiliki akurasi yang cukup tinggi. Namun, untuk memperoleh akurasi yang tinggi diperlukan beberapa optimasi yang bisa dilakukan pada SVM seperti memilih proses preprocessing yang tepat, pemilihan kernel yang tepat, dan pemilihan fitur ekstraksi yang tepat. Dalam beberapa penelitian klasifikasi teks, SVM juga cenderung sangat akurat, terutama ketika langkah-langkah optimasi tersebut diterapkan dengan baik.

Berdasarkan uraian diatas, peneliti tertarik untuk melakukan penelitian terkait perbandingan hasil klasifikasi antara *framework* DISC dan OCEAN dalam analisis kepribadian seseorang menggunakan SVM. Dengan menggunakan SVM, penelitian ini akan melakukan analisis terhadap pemilihan tahapan preprocessing, pemilihan feature extraction, pembagian dataset, dan pemilihan kernel dalam meningkatkan kinerja metode yang diusulkan dengan memperhatikan akurasi yang dihasilkan. Hasil akhir akurasi terbaik yang dihasilkan dari penelitian ini diharapkan dapat membantu dalam klasifikasi kepribadian seseorang berdasarkan framework DISC dan OCEAN dengan tepat.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang sudah dipaparkan diatas, maka rumusan masalah pada penelitian ini sebagai berikut:

a. Bagaimana pengaruh dari penerapan *preprocessing*, pembagian data pemilihan *feature extraction*, dan kernel SVM terhadap akurasi dalam menganalisis kepribadian seseorang berdasarkan tweet dalam Bahasa Indonesia menggunakan framework DISC dan OCEAN?
b. Berapa tingkat akurasi yang dihasilkan dari setiap skenario yang diusulkan dalam menganalisis kepribadian seseorang berdasarkan tweet dalam Bahasa Indonesia pada masing-masing framework DISC dan OCEAN?
c. Framework manakah yang memiliki tingkat akurasi terbaik antara DISC dan OCEAN dalam menganalisis kepribadian seseorang berdasarkan tweet mereka dalam Bahasa Indonesia?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini meliputi:

a. Dataset yang digunakan diambil dari tweet pribadi 52 users dalam sosial media Twitter dengan jumlah minimal tweet setiap user adalah 100 tweet
b. Fitur ekstraksi yang digunakan menggunakan 3 metode pembobotan yaitu TF-IDF, W-IDF, dan N-Gram
c. Algoritma klasifikasi yang digunakan yaitu SVM dan Kernel SVM
d. Pengujian kinerja model yang diusulkan dievaluasi menggunakan Confusion Matrix Multi-Class.
e. Indikator yang dipakai untuk membandingkan hasil performa adalah nilai accuracy, recall, dan F1-Score

f. Skenario yang digunakan untuk mendapatkan akurasi pada penelitian ini adalah penggunaan *stopword removal* pada *preprocessing* data, pembagian dataset, pemilihan *feature extraction*, dan pemilihan kernel SVM

1.4. Tujuan Penelitian

Adapun maksud dan tujuan yang ingin dicapai dari penelitian ini adalah:

a. Mengetahui tingkat akurasi dari masing-masing framework DISC dan OCEAN dalam menganalisis kepribadian seseorang berdasarkan tweet mereka dalam Bahasa Indonesia
b. Mengetahui skenario manakah yang memiliki tingkat akurasi terbaik dalam menganalisis kepribadian seseorang berdasarkan tweet mereka dalam Bahasa Indonesia
c. Mengetahui framework manakah yang memiliki tingkat akurasi terbaik antara DISC dan OCEAN dalam menganalisis kepribadian seseorang berdasarkan tweet mereka dalam Bahasa Indonesia

1.5. Manfaat Penelitian

Manfaat yang ingin dicapai dari penelitian ini adalah:

a. Dapat mengetahui tingkat akurasi dari masing-masing framework DISC dan OCEAN dalam menganalisis kepribadian seseorang berdasarkan tweet mereka dalam Bahasa Indonesia

b. Dapat mengetahui framework manakah yang memiliki tingkat akurasi terbaik antara DISC dan OCEAN dalam menganalisis kepribadian seseorang berdasarkan tweet mereka dalam Bahasa Indonesia
c. Dapat menjadi pedoman pengembangan sistem untuk menganalisis kepribadian seseorang.
d. Dapat digunakan oleh HR dalam melihat kepribadian calon karyawan melalui sosial media mereka.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Analisis kepribadian semakin berkembang, ditandai dengan banyaknya riset tentang penelitian tersebut. Pada tahun 2019, Artissa dkk melakukan penelitian dengan menggunakan metode analisis kepribadian Big-Five OCEAN (*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, *Neuroticism*). Dataset yang digunakan dalam penelitian ini diambil dari seluruh hasil postingan pada 250 akun sosial media Facebook, dengan menggunakan Bahasa Inggris. Sehingga jumlah total dataset yang digunakan ialah 1000 postingan. Pada penelitian ini Artissa dkk menambahkan metode klasifikasi Multinomial Naive Bayes. Hasil akurasi tertinggi yang didapat pada penelitian ini mencapai 59.9% (Artissa et al., 2019)

Selanjutnya pada tahun 2020, Ema Utami dkk juga melakukan panelitian analisis kepribadian. Namun berbeda dengan penelitian Artissa sebelumnya, kali ini Ema Utami menggunakan metode analisis kepribadian DISC (*Dominance*, *Influence*, *Steadiness*, *Compliance*). Dengan menggunakan teknik scraping, Ema Utami mengambil seluruh hasil tweet Bahasa Indonesia dari 120 akun Twitter yang kemudian dijadikan sebagai dataset. Pada penelitian ini juga menambahkan metode klasifikasi KNN dan Naive Bayes untuk meningkatkan akurasi pengenalan kepribadian. Hasil akurasi dari masing-masing klasifikasi tersebut ialah KNN dengan 28.33% dan Naive Bayes 34.16% (Utami et al., 2020).

Selanjutnya pada tahun yang sama, Hermawan dkk juga melakukan penelitian analisis kepribadian dengan menggunakan sosial media Twitter sebagai datasetnya. Dengan menggunakan teknik scraping, Hermawan dkk menggunakan 9044 tweet yang hanya berbahasa Indonesia sebagai dataset. Penelitian ini membagi persentase 70% sebagai training dan 30% sebagai testing. Dengan menambahkan metode Naive Bayes, akurasi tertinggi dari penelitian ini ialah 76.19% (Setiawan & Wafi, 2020).

Penelitian Analisis Kepribadian kembali dilakukan oleh Ninda dkk pada tahun 2020. Pada penelitian ini menggunakan metode analisis kepribadian yang sama dengan Artissa, yaitu Big-five OCEAN. Penelitian ini menggunakan postingan facebook sebagai datasetnya, dengan mengambil seluruh hasil postingan dari 170 akun Facebook yang berbahasa Inggris. Metode klasifikasi SVM juga diterapkan pada penelitian ini, dan hasil akurasi yang didapatkan 87.5% (N. A. Utami et al., 2020)

Penelitian Analisis kepribadian tidak hanya populer di Bahasa Inggris. Pada tahun 2021, Mervat dkk menggunakan dataset Bahasa Rumania yang diambil dari 10 akun sosial media Facebook, dengan total postingan 2000 post. Pada penelitian ini menggunakan metode analisis kepribadian DISC. Hasil akurasi yang dihasilkan pada penelitian ini terbilang cukup tinggi jika dibandingkan dengan hasil akurasi penelitian yang menggunakan Bahasa Indonesia, yaitu 90% (Cernian et al., 2021).

Selanjutnya pada tahun 2022, Mervat juga melakukan penelitian analisis kepribadian. Dengan memanfaatkan metode analisis Big-Five Ocean dan Sosial media facebook, dataset yang digunakan sejumlah 27.182, dengan postingan hanya

bahasa Inggris. Dalam penelitiannya ini, Mervat menambahkan algoritma *machine learning* dan membandingkan masing-masing metode, yaitu Linear SVC, Logistic Regression, Multinomial NB, dan Random Forest Classifier. Hasil akurasi tertinggi pada penelitian ini didapatkan saat menggunakan algoritma Linear SVC dengan nilai akurasi 89.37% (Ragab Bakry et al., 2022).

2.2. Keaslian Penelitian

Tabel 2.1. Matriks Literatur Review dan Posisi Penelitian
Perbandingan Framework Disc Dan Ocean Untuk Analisis Profil Pada Postingan Media Sosial dalam Bahasa Indonesia

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Personality Classification based on Facebook status text using Multinomial Naive Bayes	Y B N D Artissa, I Asror, S A Faraby (International Conferences on Data and Information Science, 2019)	Penelitian ini bertujuan untuk menganalisis kepribadian seseorang dengan menggunakan data sosial media Facebook. Sebelum memasuki tahap klasifikasi menggunakan MNB, pada penelitian ini juga melakukan proses pengurangan jumlah variasi kata dengan menggunakan proses stemming pada tahapan *preprocessing*.	Berdasarkan hasil dan analisis yang telah dilakukan dalam penelitian ini dapat disimpulkan bahwa akurasi rata rata proses klasifikasi menggunakan stemming pada *preprocessing* menghasilkan akurasi paling besar yaitu 59.9% karena proses stemming mengubah bentuk menjadi basis berdasarkan algoritma stemming yang digunakan.	Pada penelitian ini menggaris bawahi bahwa pengurangan jumlah variasi kata dapat meningkatkan akurasi sistem.	Pada penelitian ini menggunakan dataset bahasa Inggris, menggunakan teori Big Five OCEAN, dan menambahkan algoritma Multinomial Naive Bayes sebagai teknik klasifikasi. Sedangkan pada penelitian yang akan dilakukan menggunakan dataset Bahasa Indonesia, menggunakan teori DISC dan Big Five OCEAN dan menggunakan algoritma SVM dan Kernel SVM sebagai teknik klasifikasinya

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
2	K-Nearest Neighbor and Naive Bayes Classifier Comparison for Individual Character Classification on Twitter	Ema Utami, Sumarni Adi, Suwanto Raharjo, Anggit Dwi Hartanto, Aminudin Noor Ichsan (IEEE Access, 2020)	Pada penelitian ini bertujuan untuk menganalisis karakter seseorang berdasarkan tweet yang mereka posting dalam bahasa Indonesia di sosial media Twitter. Dan menggunakan algoritma klasifikasi NBC dan KNN karena sudah umum digunakan	Berdasarkan hasil pengujian penelitian ini, dapat disimpulkan bahwa algoritma NBC dan KNN mampu mengklasifikasikan profil akun Twitter ke dalam teori DISC meskipun dengan akurasi yang cukup rendah. Hasil akurasi terbaik didapatkan oleh klasifikasi NBC dengan tingkat akurasi 34.16% dan KNN 28.33%	Pada penelitian ini menyarankan untuk menggunakan algoritma klasifikasi lain untuk dijadikan bahan pembanding agar dapat mengetahui algoritma terbaik untuk mengklasifikasikan karakter individu menggunakan metode DISC. Selain itu, pada penelitian ini menggunakan pembobotan TF-IDF, sehingga perlu dicoba menggunakan pembobotan lain untuk mengetahui metode pembobotan mana yang paling optimal untuk digunakan pada masing-masing algoritma.	Pada penelitian ini hanya menggunakan pembobotan TF-IDF dan kemudian dilakukan klasifikasi menggunakan NBC dan KNN. Sedangkan pada penelitian yang akan dilakukan menggunakan 3 metode pembobotan yaitu TF-IDF, W-IDF, dan N-Gram, kemudian masing-masing hasil pembobotan dilakukan klasifikasi menggunakan SVM dan Kernel SVM.

Tabel 2.1. Lanjutan

No	Judul	Pencliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3	Classification of Personality Type Based on Twitter Data Using Machine Learning Techniques	Hermawan Setiawan, Achmad Abdul Wafi (IEEE Access, 2020)	Penelitian ini membangun model prediksi untuk mengklasifikasikan kepribadian seseorang berdasarkan data Twitter menjadi 4 kelas menurut tipe kepribadian DISC menggunakan algoritma Naive Bayes.	Berdasarkan analisis dan pengujian yang telah dilakukan, dapat diambil kesimpulan bahwa model prediktif untuk mengklasifikasikan tipe kepribadian DISC menggunakan metode Naive Bayes berhasil dilakukan. Model prediksi yang telah dibangun memiliki tingkat akurasi sebesar 76.19%. Nilai tersebut diperoleh dengan membandingkan hasil klasifikasi tipe kepribadian DISC dari model prediksi dan hasil klasifikasi tipe kepribadian DISC dari para ahli atau pakar.	Pada penelitian ini menyarankan untuk kedepannya menambah jumlah dataset khususnya pada training, dan mengimplementasikan tahap *preprocessing* lainnya untuk meningkatkan akurasi prediksi. Menggunakan algoritma klasifikasi lain untuk dapat melihat perbandingan akurasi prediksi.	Pada penelitian ini menggunakan 4 tahap *preprocessing*, yaitu case folding, stopwords removal, stemming, dan tokenization. Sedangkan pada penelitian yang akan dilakukan menambahkan normalization pada tahap *preprocessing*, yang berfungsi untuk mengubah kata dalam tweet yang disingkat menjadi kata baku.

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
4	Personality Classification of Facebook Users According to Big Five Personality Using SVM Method	Ninda Anggoro Utami, Warih Maharani, Imelda Atastina (Elsevier, 2020)	Pada penelitian ini bertujuan untuk membangun sebuah model SVM yang dapat menemukan kepribadian pengguna Facebook sesuai dengan aktivitas yang mereka lakukan di Facebook. sehingga kedepannya hasil penelitian ini dapat digunakan oleh perusahaan untuk mencari sumber daya manusia terbaik sesuai dengan bidang yang dibutuhkan oleh perusahaan.	Berdasarkan hasil yang diperoleh dari pengujian yang telah dilakukan parameter SVM terbaik yang didapatkan pada penelitian ini menggunakan Radial Basis Function (RBF) dengan nilai akurasi 87.5%	Disarankan untuk menambah jumlah dataset baru	Pada penelitian ini berhasil mendapatkan parameter terbaik dari klasifikasi SVM, sehingga parameter ini dapat menjadi acuan peneliti dalam menggunakan klasifikasi SVM.

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5	Fostering Cyber-Physical Social Systems through an Ontological Approach to Personality Classification Based on Social Media Posts	Alexandra Cernian, Nicoleta Vasile, Ioan Stefan Sacala (MDPI Journal, 2021)	Penelitian ini bertujuan untuk menganalisis postingan media sosial seseorang yaitu Facebook maupun Twitter dan akan mengidentifikasi tipe kepribadian yang dimiliki pengguna berdasarkan teori DISC.	Penelitian ini menggunakan dataset Bahasa Rumania yang diambil dari 10 akun sosial media Facebook dengan total postingan 2000 post. Pada penelitian ini menggunakan metode analisis kepribadian DISC. Hasil akurasi yang dihasilkan pada penelitian ini terbilang cukup tinggi jika dibandingkan dengan hasil akurasi penelitian yang menggunakan Bahasa Indonesia, yaitu 90%	Pada penelitian ini menyarankan untuk menambahkan jumlah dataset yang digunakan mengingat jumlah dataset yang digunakan pada penelitian ini hanya 2000 data	Dataset yang digunakan pada penelitian ini masih tergolong kurang, karena hanya 2000 data, sedangkan pada penelitian yang dilakukan akan menggunakan 5000 data tweet Bahasa Indonesia.
6	Personality Classification Model of Social Network Profiles based on their Activities and Contents	Mervat Regab Bakry, Mona Mohamed Nasr, Fahad	Penelitian ini bertujuan menganalisis kepribadian dengan metode	Hasil dari penelitian ini membuktikan keberhasilan mengidentifikasi kepribadian pengguna	Penelitian ini menyarankan untuk kedepannya menggunakan deep learning untuk dapat	Pada penelitian ini menggunakan dataset Bahasa Inggris, dan menggunakan teori kepribadian OCEAN

Tabel 2.1. Lanjutan

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		Kamal Alaheref (IJACSA, 2022)	analisis Big-Five Ocean dan Sosial media facebook, dataset yang digunakan sejumlah 27.182, dengan postingan hanya bahasa Inggris. Dengan membandingkan masing-masing metode Linear SVC, Logistic Regression, Multinomial NB, dan Random Forest Classifier.	berdasarkan postingan mereka di sosial media Facebook. Penelitian ini membandingkan metode Linear SVC, Logistic Regression, Multinomial NB, dan Random Forest Classifier. Hasil akurasi terbaik didapatkan Linear SVC dengan akurasi 89.37%	meningkatkan akurasi sistem pengenalan.	sedangkan pada penelitian yang akan dilakukan menggunakan datset tweet Bahasa Indonesia dengan teori kepribadian DISC dan Big Five OCEAN

2.3. Landasan Teori

Terdapat beberapa landasan teori yang dibutuhkan dalam penelitian ini, mulai dari landasan teori tentang kepribadian, text *preprocessing*, DISC, Big-five OCEAN.

2.3.1 Kepribadian

Kepribadian *(personality)* menurut Darlega, Winstead & Jones ialah "sistem yang relatif stabil mengenai karakteristik individu yang bersifat internal, yang berkontribusi terhadap pikiran, perasaan, dan tingkah laku yang konsisten" (Syamsu Yusuf et al., 2008). Sedangkan menurut Alwisol, kepribadian adalah sifat dasar yang dimiliki oleh seseorang yang bisa membedakannya dengan orang lain, dimana Kepribadian meliputi keseluruhan fikiran, tingkah laku, perasaan, kesadaran dan ketidaksadaran (Alwisol et al., 2009).

Kemudian Woodworth mengemukakan pendapatnya tentang kepribadian bahwa setiap perbuatan seseorang itu diwarnai oleh kepribadiannya. Baginya, "kepribadian bukanlah suatu substansi melainkan gejalanya dan suatu gaya hidup. Kepribadian tidaklah menunjukkan jenis suatu aktivitas, seperti berbicara, mengingat, berfikir, atau bercinta, tetapi seseorang individu dapat menampakkan kepribadiannya dalam cara-cara ia melakukan aktifitas-aktifitas tersebut (Patty et al., 1982).

Berdasarkan paparan diatas dapat diambil kesimpulan, bahwa kepribadian *(personality)* yaitu suatu ciri dari seseorang yang dapat mencerminkan perilaku, pemikiran, dan emosinya yang dapat membedakannya dengan orang lain dalam menghadapi dunianya.

2.3.2 Text *Preprocessing*

Text Processing merupakan Teknik untuk pengolahan data mentah atau data yang tidak terstruktur menjadi data yang terstruktur (Tri Jaka, 2015). Teknik text processing sangat dibutuhkan saat menggunakan metode crawling pada pengumpulan data. Karena data yang didapatkan masih tidak terstruktur dan tidak baku. Terdapat beberapa Langkah yang umum digunakan untuk melakukan text *preprocessing*, namun pada penelitian ini mengadopsi tahapan *preprocessing* berdasarkan penelitian yang telah dilakukan oleh Khairunnisa, dkk (Khairunnisa, 2021) karena telah berhasil menjabarkan fungsi dari masing-masing tahapan *preprocessing*. Berikut merupakan tahapan *preprocessing* yang digunakan:

1. Case Folding

   Case Folding merupakan tahapan yang berfungsi untuk mengubah semua huruf menjadi huruf kecil, kemudian case folding juga mampu menghapus tanda baca. Tahapan ini bertujuan untuk mempermudah dalam pencarian kata (Setiawan & Wafi, 2020).

2. Tokenizing

   Tokenizing merupakan tahapan yang berfungsi untuk membagi teks yang dapat berupa kalimat, paragraph atau dokumen menjadi bagian-bagian tertentu atau yang biasa disebut menjadi suatu token-token (Setiawan & Wafi, 2020).

3. *Stopword removal*

   Stopword removal merupakan tahapan berfungsi untuk membuang kata-kata yang tidak penting atau yang tidak memiliki arti (Artissa et al., 2019).

4. Normalization

   Normalization merupakan tahapan yang berfungsi untuk menormalkan sebuah kata atau kalimat. Dimana pada tahapan ini mengubah kata yang disingkat menjadi kata yang baku.

5. Stemming

   Stemming merupakan tahapan yang berfungsi untuk memberikan pemetaan pada sebuah kata-kata yang akan diubah menjadi kata dasar. Dalam bahasa Indonesia, untuk membuat kata menjadi kata dasar harus dihilangkan Prefiks, Sufiks dan Konfiks (Novák et al., 2021). Prefiks merupakan imbuhan yang berada di awal kata, Sufiks adalah imbuhan yang berada di akhiran kata, sedangkan konfiks merupakan gabungan antara Prefiks dan Sufiks (Artissa et al., 2019)

2.3.3 DISC

Pada tahun 1970, psikolog Amerika John G. Geier mengembangkan model berdasarkan penelitian berdasarkan perilaku seseorang, dimulai dari studi Moulton W. Marston— "Emotions of normal people". Psikolog Geier memaparkan bahwa terdapat empat tipe dasar kepribadian yang ditemukan pada setiap manusia, dalam proporsi yang berbeda. Kepribadian tersebut dapat dilihat pada gambar 2.1:

Gambar 2.1. The 4 Behavioral Types of the DISC Model (Somatdie et al., 2019)

1. *Dominance* (Dominan)

   Orang dengan sifat *dominance* (kuasa), suka akan tantangan dan hal bidang baru serta mengambil otoritas. Mereka akan merasa nyaman pada lingkungan penuh kekuasaan dan wewenang dengan kesempatan untuk meningkatkan prestasi individu. Orang dominan suka bertanggung jawab dan tidak ingin berada di bawah kendali orang lain dan paling nyaman. Orang dominan selalu memunculkan ide besar, sehingga sering memegang kendali dan membuat keputusan yang cepat. Akan tetapi orang dominan sering melewatkan detail, kurang berkomitmen, dan sering tidak memperdulikan penilai dan perasaan orang lain. Sehingga orang dengan *dominance* kuat, membutuhkan orang lain sebagai penasehat, menghitung resiko dan fakta penelitian. Biasanya orang dengan kecenderungan dominan akan mampu mengemban tugas sebagai pimpinan, misalnya CEO atau yang lainnya (Somatdie et al., 2019) (Cernian et al., 2021).

2. *Influence* (Mempengaruhi)

Individu dengan sifat *influence* lebih suka berada dan bekerja sama dengan orang lain. Mereka sangat senang bergaul dengan komunitas atau suatu kelompok untuk memperbanyak relasi. Orang *influence* menikmati berhubungan dengan orang lain dan membuat kesan baik, berbicara lantang dan menciptakan lingkungan yang positif dan antusias. Orang dengan sifat *influence* lebih suka dalam bidang pembinaan dan konseling. Sayangnya, orang dengan tipe ini memiliki perasaan yang halus dan tidak berkonsentrasi dengan tugas yang sekarang dihadapi. Mereka lebih cocok sebagai penasehat. Mereka membutuhkan orang lain untuk mencari fakta, berkomunikasi lengkap, menghormati ketulusan, memberi penghargaan atas hal-hal kecil, melakukan pendekatan secara logis. (Somatdie et al., 2019) (Cernian et al., 2021)

3. *Steadiness* (Stabil)

Seseorang dengan *steadiness* dikenal dengan konsistensinya melakukan sesuatu sampai berhasil/selesai, tidak berorientasi pada kecepatan tetapi konsistensi. Kontribusi paling positif untuk orang *stediness* ini adalah menjadi pendengar sejati, sabar, suka membantu dan pandai mengendalikan keadaan. Individu dengan kategori ini selalu focus pada bekerja sama dengan orang lain untuk menyelesaikan tugas mereka. Kelemahannya adalah tidak memiliki motivasi diri yang kuat, sehingga cepat berubah dsan mudah dipengaruhi orang lain dan enggan membuat keputusan. Tipe ini memerlukan orang yang dapat menekan mereka dan

membantu memprioritaskan tugas, pekerjaan dan memiliki fleksibilitas tinggi dalam prosedur kerjanya. Dalam rangka memiliki efektifitas yang optimal, individu ini perlu diberitahu secara mendalam tentang perubahan yang akan datang sesegera mungkin agar dapat menyesuaikan. Orang *steady* akan nyaman jika berada pada lingkungan yang minimal konflik, berorientasi pada tugas kelompok, memiliki penghargaan yang tulus dan percaya pada kemampuan orang lain (Cernian et al., 2021).

4. Compliance (Pemikir)

Seseorang dengan sifat *compliance* adalah sangat teliti dan suka berpikir rumit. Orang *compliance* biasanya teguh dalam pendirian dan pilihannya. Sifat positif yang dapat diambil dari *compliance* adalah teliti, berfikir kritis, menggunakan pendekatan secara halus dan analitis, memiliki rencana yang matang, dapat menyelesaikan masalah dengan baik, *profesional*, diplomatis dan punya loyalitas tinggi. Kelemahannya adalah cenderung meragu dan terkesan lambat dalam pengambilan keputusan karena terlalu teliti. Maka, pekerjaan yang dilakukan akan terproses dengan lambat, pendendam dan terlalu kritis. Orang yang sangat teliti, akan membutuhkan orang lain untuk berkompromi dan mengambil keputusan yang cepat. Sehingga orang *compliance* biasanya bekerja di bawah komando orang *dominance* sebagai Asisten Manajer atau yang lainnya (Cernian et al., 2021).

2.3.4 Big Five OCEAN

Big five merupakan salah satu pendekatan yang digunakan untuk melihat kepribadian seseorang berdasarkan lima dimensi yang telah dibentuk dengan menggunakan analisis faktor. Big five sering digambarkan sebagai kerangka yang bersifat universal untuk mengukur kepribadian individu secara komprehensif (Nurhayati et al., 2020). Kelima sifat dasar tersebut ialah:

1. *Extraversion*

   Seseorang dengan sifat *extraversion* cenderung percaya diri, dominan, aktif, dan menunjukkan emosi yang positif, selain itu juga dikaitkan dengan kecenderungan untuk bersikap optimis. Seorang *extraversion* lebih cepat bergaul sehingga mudah untuk berteman dengan orang lain, dan mudah termotivasi oleh perubahan (N. A. Utami et al., 2020); (Ragab Bakry et al., 2022)

2. *Agreeableness*

   Seseorang dengan sifat *Agreeableness* mampu beradaptasi sosial yang baik mengindikasikan individu yang ramah, memiliki kepribadian yang selalu mengalah, menghindari sebuah konflik dan memiliki kecenderungan untuk mengikuti orang lain. Seseorang yang memiliki *Agreeableness* yang tinggi digambarkan sebagai seseorang yang memiliki *value* suka membantu, forgiving, dan penyayang (N. A. Utami et al., 2020); (Ragab Bakry et al., 2022)

3. *Neuroticism*

Seseorang dengan sifat *Neuroticism* memiliki emosi yang negative seperti, rasa khawatir, cemas, rasa tidak aman dan labil. Individu yang memiliki nilai yang tinggi dalam dimensi ini kepribadiannya mudah mengalami kecemasan, rasa marah, depresi. (N. A. Utami et al., 2020); (Ragab Bakry et al., 2022)

4. *Conscientiousness*

Seseorang dengan sifat *Conscientiousness* pada umumnya berhati-hati, dapat diandalkan, teratur dan bertanggung jawab (N. A. Utami et al., 2020); (Ragab Bakry et al., 2022)

5. *Openness*

Seseorang dengan sifat *Openness* senang dengan informasi baru, dan juga mengacu pada bagaimana individu-individu bersedia melakukan penyesuaian pada suatu ide atau situasi yang baru, mudah bertoleransi, memiliki kapasitas untuk menyerap informasi, focus dan kreatif dan artistic (N. A. Utami et al., 2020); (Ragab Bakry et al., 2022)

2.3.5 *Feature extraction*

Fitur ekstraksi dalam *Natural Language Processing* NLP merupakan proses mengubah data teks menjadi representasi numerik yang dapat digunakan oleh algoritma pembelajaran mesin. Tujuan adanya fitur ekstraksi ialah mengubah teks menjadi representasi yang lebih terstruktur agar dapat dimengerti oleh model

pembelajaran mesin (Liang et al., 2017). Terdapat beberapa metode fitur ekstraksi dalam NLP, seperti:

1. TF-IDF (Term Frequency-Inverse Document Frequency)

   Metode ini menghitung dua faktor untuk setiap kata dalam sebuah dokumen. Dimana frekuensi kemunculan kata dalam dokumen disebut (TF) dan sebaliknya frekuensi kemunculan kata di seluruh dokumen dalam korpus disebut (IDF). Hasilnya adalah vektor yang merepresentasikan bobot kata-kata dalam dokumen. Setiap elemen vektor merepresentasikan bobot TF-IDF dari kata tersebut (Berliana & Shaufiah, 2018). Rumus persamaan dari TF-IDF ialah sebagai berikut:

$$tf_{td}idf_t = tf_{td} * log\ log\ (\frac{N}{df_t})$$

   Dimana
   tf_{td} x idf : Bobot total dari kata t
   tf_{td} : Jumlah kemunculan kata t dalam suatu dokumen
   N : Total dokumen
   df_t : Jumlah dari seluruh dokumen yang mengandung kata t

   Semakin sering sebuah fitur muncul dalam sebuah teks, maka semakin besar pula bobot yang akan didapat, dimana ini berarti semakin penting fitur tersebut (Berliana & Shaufiah, 2018).

2. W-IDF (Word Importance based on Document Frequency)

   W-IDF merupakan modifikasi dari metode TF-IDF yang mempertimbangkan tingkat pentingnya kata berdasarkan frekuensi dokumen dimana kata tersebut muncul. W-IDF memberikan bobot yang lebih tinggi untuk kata-kata yang muncul dalam sedikit dokumen,

menunjukkan bahwa kata-kata tersebut lebih unik dan dapat memberikan informasi yang lebih berharga dalam representasi teks.

3. N-Gram

N-gram merupakan metode membagi teks menjadi bagian-bagian yang lebih kecil berdasarkan urutan n kata yang berdekatan. N-gram digunakan untuk mempertahankan informasi urutan kata dalam teks. Teknik N-gram didasarkan pada pemisahan teks menjadi *string* dengan panjang n mulai dari posisi tertentu dalam suatu teks (Miftahuddin et al., 2016) . Posisi N-G=gram berikutnya dihitung dari posisi yang sebenarnya bergeser sesuai dengan *offset* yang diberikan. N-gram untuk setiap string dihitung dan kemudian dibandingkan satu per satu. N-gram dapat berupa unigram (n=1), bigram (n=2), trigram (n=3) dan seterusnya (Miftahuddin et al., 2016).

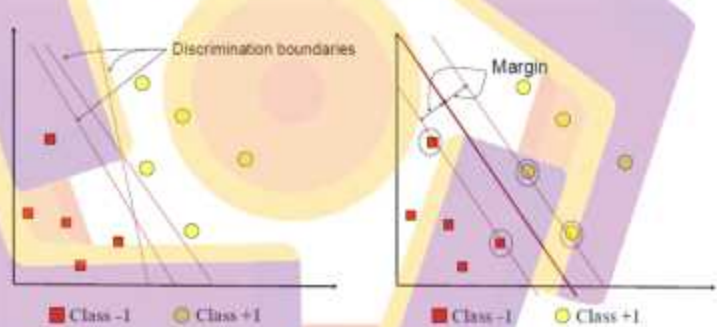
2.3.6 SVM

Teknik klasifikasi dalam pemrosesan NLP adalah metode untuk mengklasifikasikan teks atau dokumen ke dalam kategori atau label yang sudah ditentukan sebelumnya. Teknik ini sangat penting dalam NLP karena memungkinkan untuk mengenali pola, tema, atau makna dari teks tersebut. Berikut ini adalah beberapa teknik klasifikasi yang umum digunakan dalam NLP:

1. SVM (Support Vector Machine)

SVM merupakan metode klasifikasi yang mencari hyperplane dengan margin terbesar. Umumnya SVM mirip dengan ANN yang termasuk dalam supervised learning yaitu mencari hyperplane dengan cara

memisahkan kumpulan data menjadi dua kelas yang berbeda. Hyperplane akan menemukan titik optimalnya ketika jaraknya tepat berada di tengah kelas yang telah dipisahkan. Inti dari cara kerja SVM adalah mencari jarak terjauh dari hyperplane yang terbagi menjadi dua kelas. Proses penyelesaian jarak terjauh diulang beberapa kali untuk menemukan hyperplane terbaik. Oleh karena itu dibutuhkan 7 optimasi pada SVM untuk mencari jarak maksimum pada hyperplane dengan kedua kelas tersebut. Ada dua bentuk optimasi yang dimaksudkan untuk menemukan hyperplane dalam SVM, yaitu SVM Primal Form dan SVM Dual Form. Gambar 2.2 merupakan gambaran SVM yang mencoba mencari hyperplane terbaik.



Gambar 2.2. Hyperplane SVM

Gambar 2.2 menjelaskan bahwa SVM mencoba mencari hyperplane gambar a menunjukkan pola dua kelas, yaitu +1 dan -1. Warna merah merupakan pola yang termasuk dalam kelas -1, sedangkan warna kuning merupakan pola yang termasuk dalam kelas +1. Masalah dengan klasifikasi ini adalah SVM mencoba menemukan hyperplane yang memisahkan kedua kelas. Batas diskriminasi adalah simbol dari garis pemisah alternatif.

Hyperplane terbaik dapat ditemukan dengan mengukur margin dan menemukan titik optimalisasi. Margin adalah jarak antara hyperplane dan pola terdekat dari masing-masing kelas tersebut. Garis paling tebal pada gambar b menunjukkan hyperplane terbaik. Pada metode SVM digunakan parameter tuning yaitu sebagai parameter value atau pengaturan parameter untuk mengidentifikasi dan memecahkan masalah yang dihadapi. Salah satu rumus formula SVM adalah sebagai berikut:

$$L(\vec{w}, b, a) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} ai(yi\{(\overrightarrow{xi}.\vec{w} + b) - 1\})$$

αi adalah Lagrange multipliers, yang bernilai nol atau positif (αi > 0). Nilai optimal dari persamaan diatas dapat dihitung dengan meminimalkan L terhadap w dan b, dan memaksimalkan L terhadap αi. Dengan memperhatikan sifat bahwa pada titik optimal gradient L=0, persamaan diatas dapat dimodifikasi sebagai maksimalisasi problem yang hanya mengandung αi saja, sebagaimana pada persamaan 7 dibawah:

Maximize:

$$\sum_{i=1}^{l} ai - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j \vec{x}_i \vec{x}_j$$

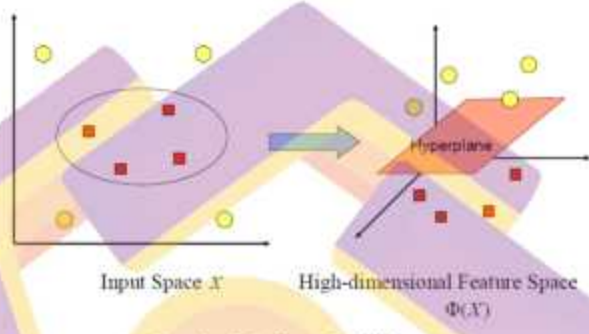
Subject to:

$$a_i \geq 0\ (i = 1,2,\ldots,l) \quad \sum_{i=1}^{l} a_i y_i = 0$$

Dari hasil perhitungan ini diperoleh αi yang kebanyakan bernilai positif. Data yang berkolerasi dengan αi yang positif inilah yang disebut sebagai support vector.

2. Metode Kernel

Secara umum dalam dunia nyata seringkali tidak bersifat linear separable, dimana biasanya kebanyakan bersifat non linear. Untuk mengatasi hal ini, SVM dimodifikasi dengan memasukkan fungsi kernel.



Gambar 2.3. Kernel SVM

Seperti yang terlihat pada Gambar 2.3, dalam SVM non linear, pertama-tama data x dipetakan oleh fungsi $\Phi(x)$ ke ruang vektor yang berdimensi lebih tinggi. Pada ruang vektor yang baru ini, hyperplane yang memisahkan kedua kelas tersebut dapat dikonstruksikan. Dapat dilihat pada gambar diatas, data pada class kuning dan data pada class merah yang berada pada input space berdimensi dua tidak dapat dipisahkan secara linear. Selanjutnya pada gambar 2b menunjukkan bahwa fungsi Φ memetakan tiap data pada input space tersebut ke ruang vektor baru yang berdimensi lebih tinggi (dimensi 3), dimana kedua class dapat dipisahkan secara linear oleh sebuah hyperplane. Tabel 2.2 merupakan jenis-jenis kernel yang umum digunakan dalam SVM sedangkan formula matematika dari keadaan ini ditunjukkan pada persamaan dibawah:

$$\theta : R^d \rightarrow R^q \quad d < q$$

Tabel 2.2. Kernel yang umum digunakan SVM

Jenis Kernel	Rumus
Polynomial	$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i . \vec{x}_j + 1)^p$
RBF	$K(\vec{x}_i, \vec{x}_j) = exp\ exp\ (-\frac{\|\vec{x}_i . \vec{x}_j\|^2}{2\sigma^2})$
Sigmoid	$K(\vec{x}_i, \vec{x}_j) = tanh\ tanh\ (a\vec{x}_i, \vec{x}_j + \beta)$

2.3.7 *Split* Data

Tahap *split*ting dataset dalam penelitian ini bertujuan untuk memastikan akurasi terbaik dan memastikan distribusi data yang lebih representatif antara data training dan data testing. Pendekatan *split*ting data efektif baik untuk masalah klasifikasi [*Split*_An_Optimal_Method_for_Data_*Split*ting] dan memberikan solusi yang lebih andal dalam pembagian data untuk pelatihan dan evaluasi model. Proses *split*ting data pada penelitian ini dilakukan pembagian dataset sebesar 4 tahapan, tahapan pertama training 90% dan testing 10%; tahapan kedua training 80% dan testing 20%; tahapan ketiga training 70% dan testing 30%; tahapan keempat training 60% dan testing 40%.

2.3.8 Evaluation Method

Tahap evaluasi dalam penelitian ini menggunakan presisi, recall, f1-score, dan akurasi sebagai tolak ukur penilaian kinerja pada sistem analisis kepribadian. Presisi memberikan persentase hasil positif yang relevan, sedangkan recall mengukur seberapa baik model mengidentifikasi semua kasus relevan. Sementara

F1-score merupakan rata-rata dari presisi dan recall dan perhitungan nilai akurasi diperoleh dengan membandingkan hasil prediksi yang ada dengan data kelas sebenarnya. Penggunaan keempat metrik ini sangat penting untuk mengevaluasi secara efektif performa model dalam klasifikasi, sehingga dapat memberikan pemahaman yang lebih mendalam terkait kelebihan dan kekurangan model yang digunakan, Persamaan rumus perhitungan masing-masing evaluasi dapat dilihat pada persamaan 8 dibawah ini:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2\,(precision\ x\ recall)}{precision + recall}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Jenis dan pendekatan penelitian ini adalah penelitian kuantitatif, serta sifat dari penelitian ini merupakan eksperimental komputasi. Penelitian ini menggunakan pendekatan kuantitatif untuk melakukan suatu eksperimen yang digunakan untuk menganalisa dan melakukan komparasi terhadap performa model SVM dan Kernel SVM dalam klasifikasi kepribadian seseorang berdasarkan tweet mereka di sosial media. Sifat dari penelitian ini dilakukan secara mandiri menggunakan metode deskriptif dan kausal. Penggunaan metode deskriptif bertujuan untuk menggambarkan secara sistematis dari data yang diperoleh kemudian dilatih dan diuji.

3.2. Metode Pengumpulan Data

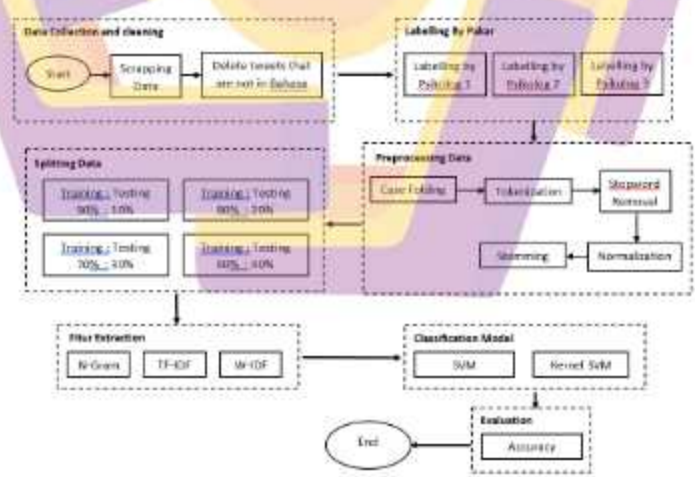
Data yang digunakan pada penelitian ini merupakan jenis data primer yang dikumpulkan langsung menggunakan teknik crawling dari sosial media Twitter. Crawling data Twitter diambil berdasarkan user yang bersedia tweetnya di crawling. User yang didapatkan pada penelitian ini sejumlah 52 users, dan data tweet masing-masing users yang diambil sejumlah 100 tweet, sehingga total dataset yang digunakan 5200 tweet.

3.3. Metode Analisis Data

Sebelum melakukan analisis data, tweet yang didapatkan dilabelkan oleh pakar terlebih dahulu. Kemudian tweet yang telah didapatkan tersebut di pre-processing, seperti mengubah data menjadi huruf kecil (case folding), menghilangkan karakter ACII, menghilangkan link, menghilangkan retweet, tokenization, serta stemming menggunakan library Sastrawi.

Data yang telah dilakukan pre-processing, kemudian dilakukan *split*ting atau pembagian data menjadi dua jenis, yaitu data training dan data testing. Setelah itu data akan diolah menggunakan *feature extraction* yang terdiri dari TF-IDF, WF-IDF, N-gram dan selanjutnya dianalisis menggunakan metode SVM untuk didapatkan nilai akurasi.

3.4. Alur Penelitian



Gambar 3.1. Alur Penelitian

Gambar 3.1 merupakan diagram alur langkah penelitian secara lengkap dan terinci termasuk di dalamnya tercermin algoritma, rute, pemodelan-pemodelan, desain, yang terkait dengan aspek perancangan sistem. Pada penelitian ini terdapat beberapa langkah utama pada penelitian seperti pada gambar 5 dengan rincian alur sebagai berikut:

1. Pengumpulan dan pembersihan data

   Pada pengumpulan data, data diambil dari sosial media twitter dengan menggunakan teknik scraping. Jumlah user twitter yang diambil sejumlah 52 user, dengan masing-masing user diambil tweet yang hanya menggunakan Bahasa Indonesia sejumlah 100 tweets. Sehingga total dataset yang digunakan berjumlah 5200 data tweet

2. Pemberian Label

   Pada proses ini pemberian label dilakukan pada setiap tweet dari masing-masing user, yang berfungsi untuk memberikan informasi kepribadian dari setiap tweet tersebut. Pemberian label dilakukan oleh 3 pakar psikolog, dimana pemberian label ini merujuk kepada teori kepribadian DISC dan Big Five OCEAN.

3. *Preprocessing* Data

   Tahapan *preprocessing* data merupakan tahapan yang digunakan untuk mengolah dan mengelola data sehingga siap untuk digunakan. Pada tahapan ini terdapat beberapa proses yang dilakukan, seperti:

i. *Case Folding*

Pada proses ini mengubah semua tweet menjadi huruf kecil, dan juga menghapus tanda baca.

ii. *Tokenization.*

Pada proses ini membagi text yang berupa kalimat menjadi sebuah token-token.

iii. *Stopword removal.*

Pada proses ini membuang kata-kata yang tidak penting atau yang tidak memiliki arti.

iv. *Normalization*

Pada proses ini menormalkan sebuah kata atau kalimat. Dimana pada tahapan ini mengubah kata yang disingkat menjadi kata yang baku.

v. *Stemming*

Stemming merupakan tahapan yang berfungsi untuk memberikan pemetaan pada sebuah kata-kata yang akan diubah menjadi kata dasar.

4. *Splir*ting data

Proses *split* data dilakukan pembagian dataset sebesar 4 tahapan, tahapan pertama training 90% dan testing 10%; tahapan kedua training 80% dan testing 20%; tahapan ketiga training 70% dan testing 30%; tahapan keempat training 60% dan testing 40%. *Split* data dilakukan pada dataset yang berjumlah 5000 data tweet.

5. *Fitur Extraction*

Tahapan *Fitur Extraction* bertujuan untuk mengubah teks menjadi representasi yang lebih terstruktur agar dapat dimengerti oleh model pembelajaran mesin. Pada penelitian ini menggunakan 3 fitur ekstraksi yang berbeda, yaitu:

i. N-Gram

N-gram merupakan metode membagi teks menjadi bagian-bagian yang lebih kecil berdasarkan urutan n kata yang berdekatan. N-gram digunakan untuk mempertahankan informasi urutan kata dalam teks.

ii. TF-IDF

Metode ini menghitung dua faktor untuk setiap kata dalam sebuah dokumen. Dimana frekuensi kemunculan kata dalam dokumen disebut (TF) dan sebaliknya frekuensi kemunculan kata di seluruh dokumen dalam korpus disebut (IDF).

iii. W-IDF

W-IDF memberikan bobot yang lebih tinggi untuk kata-kata yang muncul dalam sedikit dokumen, menunjukkan bahwa kata-kata tersebut lebih unik dan dapat memberikan informasi yang lebih berharga dalam representasi teks.

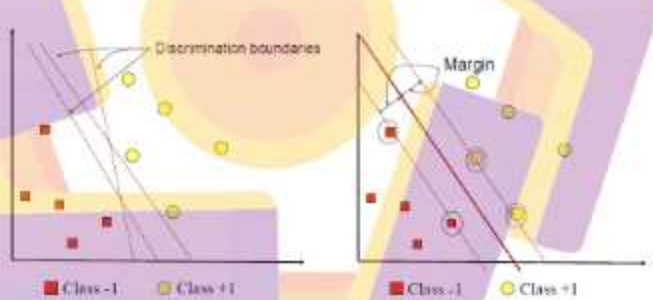
6. Classification Model

Klasifikasi merupakan metode untuk mengklasifikasikan teks atau dokumen ke dalam kategori atau label yang sudah ditentukan sebelumnya.

Metode ini memungkinkan untuk mengenali pola, tema, atau makna dari teks tersebut. Pada penelitian ini menggunakan 2 metode klasifikasi, yaitu:

i. SVM

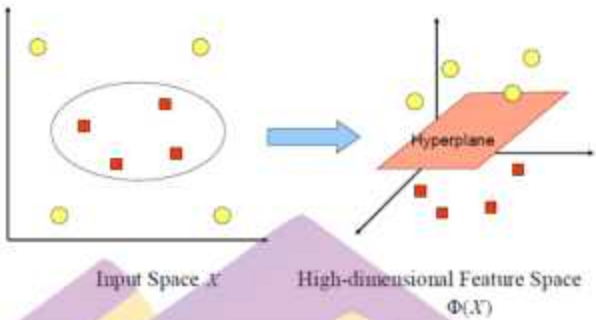
Cara kerja SVM adalah mencari jarak terjauh dari hyperplane yang terbagi menjadi dua kelas. Proses penyelesaian jarak terjauh diulang beberapa kali untuk menemukan hyperplane terbaik. Oleh karena itu dibutuhkan 7 optimasi pada SVM untuk mencari jarak maksimum pada hyperplane dengan kedua kelas tersebut. Ada dua bentuk optimasi yang dimaksudkan untuk menemukan hyperplane dalam SVM, yaitu SVM Primal Form dan SVM Dual Form. Gambar 3.2 merupakan gambaran SVM yang mencoba mencari hyperplane terbaik.



Gambar 3.2. Hyperplane SVM

ii. Kernel SVM

Secara umum dalam dunia nyata seringkali tidak bersifat linear separable, dimana biasanya kebanyakan bersifat non-linear. Untuk mengatasi hal ini, SVM dimodifikasi dengan memasukkan fungsi kernel.



Gambar 3.3. Kernel SVM

Dalam SVM non linear seperti pada gambar 3.3, pertama-tama data x dipetakan oleh fungsi Φ(x) ke ruang vektor yang berdimensi lebih tinggi. Pada ruang vektor yang baru ini, hyperplane yang memisahkan kedua kelas tersebut dapat dikonstruksikan. Dapat dilihat pada gambar diatas, data pada class kuning dan data pada class merah yang berada pada input space berdimensi dua tidak dapat dipisahkan secara linear. Selanjutnya pada gambar 2b menunjukkan bahwa fungsi Φ memetakan tiap data pada input space tersebut ke ruang vektor baru yang berdimensi lebih tinggi (dimensi 3), dimana kedua class dapat dipisahkan secara linear oleh sebuah hyperplane.

7. Evaluasi dan Pembahasan

Tahapan ini digunakan untuk melakukan evaluasi terhadap setiap pelatihan menggunakan skenario yang telah disebutkan pada alur penelitian. Terdapat dua hasil evaluasi yang dilakukan, yaitu mengetahui tingkat performa dari masing-masing skenario yang digunakan, selanjutnya membandingkan hasil akurasi dari masing-masing skenario.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Pengumpulan Data

Data yang digunakan pada penelitian ini merupakan kumpulan tweet berbahasa Indonesia yang dikumpulkan langsung menggunakan teknik crawling dari sosial media twitter. Crawling data diambil berdasarkan user yang bersedia tweetnya di crawling. Proses crawling tweet ini dapat dilihat pada Gambar 4.1

Gambar 4.1. Proses Crawling Dataset

Berdasarkan Gambar 4.1, proses crawling dilakukan dengan *TwitterSearchScraper* dari *snscrape* dengan parameter yang digunakan adalah username karena akan digunakan untuk mencari tweet dari pengguna tertentu. Melalui proses perulangan, tweet yang ditemukan akan ditambahkan ke dalam tweets_list1 hingga mencapai batas maksimal yang ditentukan yaitu 300 tweet. Setiap tweet yang dikumpulkan akan mencakup tanggal tweet, ID tweet, isi tweet, dan nama pengguna. Setelah semua tweet dikumpulkan, daftar tersebut dikonversi

menjadi DataFrame pandas dengan kolom-kolom yang telah ditentukan, yaitu 'Datetime', 'Tweet Id', 'Text', dan 'Username'.

Jumlah user yang di-*crawling* berjumlah 52 user dengan masing-masing user diambil tweet yang hanya menggunakan Bahasa Indonesia sejumlah 300 tweets. Namun, karena yang dibutuhkan merupakan tweet yang dapat menggambarkan kepribadian user, tweet-tweet tersebut selanjutnya dilakukan beberapa tindakan seperti menghapus tweet yang hanya berupa emoji, menghapus tweet yang terlalu banyak menggunakan bahasa selain bahasa indonesia, dan tweet yang hanya hasil retweet dari postingan lain. Sehingga dari 52 user tersebut akhirnya tiap-tiap user hanya terdiri dari 100 tweet. Sehingga total dataset yang digunakan berjumlah 5200 data tweet.

4.2. Labelling dan *Preprocessing* Data

Pada proses ini pemberian label dilakukan pada setiap tweet dari masing-masing user, yang berfungsi untuk memberikan informasi kepribadian dari setiap tweet tersebut. Pemberian label dilakukan oleh 3 pakar psikolog, dimana pemberian label ini merujuk kepada teori kepribadian DISC dan Big Five OCEAN. Dari hasil pelabelan ketiga pakar tersebut, label akhir yang digunakan adalah label yang memiliki kesamaan minimal antara dua pakar. Detail jumlah dataset berdasarkan framework DISC dilihat pada Tabel 4.1, sedangkan pada OCEAN dapat dilihat pada Tabel 4.2

Tabel 4.1. Detail Persebaran Dataset Berdasarkan Framework DISC

Kelas	Jumlah per Tweet	Jumlah per akun
Dominance	318	1
Influence	2527	29
Steadiness	1734	18
Conscientiousness	621	4

Tabel 4.2. Pelabelan Dataset Berdasarkan Framework OCEAN

Kepribadian	Jumlah per Tweet	Jumlah per akun
Openness	1948	31
Extraversion	237	2
Conscientiousness	845	5
Agreeableness	1029	6
Neuroticism	1141	8

Detail persebaran dataset untuk pelabelan akun pada framework DISC berdasarkan hasil perhitungan kepribadian yang sering muncul dapat dilihat pada Tabel 4.3, sedangkan pelabelan akun pada framework OCEAN dapat dilihat pada Tabel 4.4. Namun, label-label ini hanya ketika data belum dibagi menjadi data training dan data testing. Ketika sudah dibagi, besar kemungkinan label yang ada akan menjadi label yang berbeda karena menyesuaikan dengan data yang diproses selama proses klasifikasi karena ada beberapa user memiliki nilai yang hampir mirip untuk dua kelas, sehingga ketika proses split data dilakukan secara random, dengan rasio yang berbeda-beda pula, hasil pelabelan akhirnya akan berbeda dengan data yang sudah ditampilkan di awal.

Tabel 4.3. Proses Pelabelan dan Jumlah Tweet untuk Framework DISC

No	User	D	I	S	C	Label
1	ekasullistiani	5	10	7	78	Conscientiousness
2	ana_ot	0	91	9	0	Influence
3	adputri_	0	96	1	3	Influence
4	alitapuspa_	0	44	56	0	Steadiness
5	_sylvialestari	15	46	29	10	Influence

Tabel 4.3. Lanjutan

6	adityosusanto_	9	65	23	3	Influence
7	Adamumemo	18	6	22	54	Conscientiousness
8	Aiuchida	0	22	40	38	Steadiness
9	anggycaa	3	23	74	0	Steadiness
10	a2lir	8	66	26	0	Influence
11	amandapht	0	94	3	3	Influence
12	amhryn	1	33	66	0	Steadiness
13	akuuuberuang	2	42	26	30	Influence
14	alfinalfinn	1	81	18	0	Influence
15	abcdenjiii	3	80	13	4	Influence
16	adityandr	0	96	1	3	Influence
17	_marfami	0	63	37	0	Influence
18	amarasyawalni	0	77	11	12	Influence
19	_avocada	0	68	31	1	Influence
20	anzjani	3	46	43	8	Influence
21	alaniafitri	1	65	31	3	Influence
22	aldimaulana48	5	38	40	17	Steadiness
23	adlfynf	0	90	0	10	Influence
24	_Chaendelier_	2	34	64	0	Steadiness
25	alfian_ay	21	42	37	0	Influence
26	alfrisadivaw	24	44	32	0	Influence
27	anissa_sani	2	11	77	10	Steadiness
28	ardianitaap	2	87	10	1	Influence
29	Adewiana14	4	17	55	24	Steadiness
30	apaanmanggil	0	34	63	3	Steadiness
31	amindwiananda	73	23	4	0	Dominance
32	arangkecap	25	13	51	11	Steadiness
33	bananaoatmealz_	2	16	16	66	Conscientiousness
34	andrisuartikaa	0	64	36	0	Influence
35	andiniayu	1	76	19	4	Influence
36	agus_trianto	0	17	48	35	Steadiness
37	jungsasha_	5	14	25	56	Conscientiousness
38	almostbeyours	17	31	39	13	Steadiness
39	apdesti_	2	92	5	1	Influence
40	afidazkyy	0	30	22	48	Influence
41	anakyangtangkas	0	85	15	0	Influence
42	akhsaraa	18	12	65	5	Steadiness

Tabel 4.3. Lanjutan

43	aisyahafr	0	34	47	19	Steadiness
44	9ita7unn	5	59	28	8	Influence
45	_nilazka	0	66	33	1	Influence
46	aliencuk	0	65	35	0	Influence
47	alifahdellaf	0	53	47	0	Influence
48	alaaini1	0	46	48	6	Steadiness
49	adorablejuneya	14	2	77	7	Steadiness
50	adityalinardi	7	37	51	5	Steadiness
51	adindahapsa	0	43	47	10	Steadiness
52	achadianrani	20	38	31	11	Influence

Tabel 4.4. Proses Pelabelan dan Jumlah Tweet untuk Framework OCEAN

No	User	O	C	E	A	N	Label
1	ckasullistiani	0	0	1	10	89	Neuroticism
2	ana_ot	50	9	2	25	14	Openness
3	adputri_	35	2	20	34	9	Openness
4	alitapuspa_	37	0	11	45	7	Agreeableness
5	_sylvialestari	26	2	37	13	22	Extraversion
6	adityosusanto_	69	2	12	10	7	Openness
7	Adamumcmo	36	1	31	20	12	Openness
8	Aiuchida	22	6	14	27	31	Neuroticism
9	anggycaa	61	0	9	4	26	Openness
10	a2lir	31	0	25	19	25	Openness
11	amandapht	16	1	27	32	24	Agreeableness
12	amhryn	20	2	29	16	33	Neuroticism
13	akuuuberuang	45	6	5	13	31	Openness
14	alfinalfinn	55	7	13	22	3	Openness
15	abcdenjiii	35	1	52	6	6	Extraversion
16	adityandr	20	0	15	46	19	Agreeableness
17	_marfami	35	3	23	11	28	Openness
18	amarasyawalni	15	1	43	3	38	Extraversion
19	_avocada	14	0	40	4	42	Neuroticism
20	anzjani	41	1	15	22	21	Openness
21	alaniafitri	53	0	9	22	16	Openness
22	aldimaulana48	46	0	6	33	15	Openness
23	adlfynf	34	11	27	16	12	Openness

Tabel 4.4. Lanjutan

24	_Chaendelier_	33	0	30	7	30	Openness
25	alfian_ay	53	3	0	40	4	Openness
26	alfrisadivaw	37	0	1	22	40	Neuroticism
27	anissa_sani	64	0	6	6	24	Openness
28	ardianitaap	32	3	16	24	25	Openness
29	Adewiana14	33	0	26	21	20	Openness
30	apaanmanggil	13	0	27	23	37	Neuroticism
31	amindwiananda	59	1	3	14	23	Openness
32	arangkecap	48	4	13	8	27	Openness
33	bananaoatmealz__	9	72	7	6	6	Conscientiousness
34	andrisuartikaa	12	0	15	35	38	Neuroticism
35	andiniayu	79	0	5	7	9	Openness
36	agus_trianto	62	0	1	35	2	Openness
37	jungsasha_	8	68	8	8	8	Conscientiousness
38	almostbeyours	43	0	4	13	40	Openness
39	apdesti_	25	4	43	12	16	Extraversion
40	afidazkyy	39	7	21	19	14	Openness
41	anakyangtangkas	82	0	7	11	0	Openness
42	akhsaraa	20	4	33	14	29	Extraversion
43	aisyahafr	39	2	9	29	21	Openness
44	9ita7unn	49	9	13	22	7	Openness
45	_nilazka	10	0	21	47	22	Agreeableness
46	aliencuk	33	0	3	37	27	Agreeableness
47	alifahdellaf	41	0	19	19	21	Openness
48	alaaini1	41	0	0	14	45	Neuroticism
49	adorablejuncya	44	4	7	23	22	Openness
50	adityalinardi	74	0	12	5	9	Openness
51	adindahapsa	33	0	9	40	18	Agreeableness
52	achadianrani	37	1	19	15	28	Openness

Selanjutnya, sampel dataset untuk hasil pelabelan berdasarkan framework DISC dapat dilihat pada Tabel 4.5, sedangkan dataset untuk pelabelan yang dilakukan berdarkan framework Big Five OCEAN dapat dilihat pada Tabel 4.6.

Tabel 4.5. Pelabelan Dataset Berdasarkan Framework DISC

Username	Tweet	Label
ekasullistiani	Pengen banget k bali tp tiket pesawatnya bikin ku menangisssssss	Influence
ekasullistiani	Dengerin lagu aja sampe nangis bombay	Influence
ekasullistiani	Pengen banget seketika jadi gila biar lupa semuanya yg bikin sakit	Conscientiousness
ekasullistiani	Ngomong salah, diem salah, nangis pun juga salah. Gimana cara saya meluapkan emosi saya, sakit saya?	Conscientiousness
ekasullistiani	Pengen marah, pengen ngeluarin semua unek2 tapi ttep gabisa. Kenapa saya harus selalu mengerti disaat hati saya sakitnya udah gak bisa diungkapkan lagi💔	Conscientiousness
ekasullistiani	Dibalik suksesnya ktt asean di labuan bajo, ada kekecewaan dari yang bertugas jaga keamanan disana. Udah capek tenaga, waktu, pikiran, dll tp semuanya gk dihargai. Cuma bisa ngeluh sesama temen yg tugas, gk bisa protes juga. Kasihan banget sumpah.	Steadiness
ekasullistiani	Mau gimana juga kalo emg dasarnya udah egois, gk sadar diri ttep gak bakal pernah bisa berubah. Berujung jadi suka bohong.	Dominance
ekasullistiani	Hp lebih penting dari pada apapun😐	Dominance

Tabel 4.6. Pelabelan Dataset Berdasarkan Framework OCEAN

Username	Tweet	Label
ekasullistiani	Pengen banget k bali tp tiket pesawatnya bikin ku menangisssssss	Neuroticism
ekasullistiani	Dengerin lagu aja sampe nangis bombay	Neuroticism
ekasullistiani	Pengen banget seketika jadi gila biar lupa semuanya yg bikin sakit	Neuroticism
ekasullistiani	Ngomong salah, diem salah, nangis pun juga salah. Gimana cara saya meluapkan emosi saya, sakit saya?	Neuroticism
ekasullistiani	Pengen marah, pengen ngeluarin semua unek2 tapi ttep gabisa. Kenapa saya harus selalu mengerti disaat hati saya sakitnya udah gak bisa diungkapkan lagi💔	Neuroticism
ekasullistiani	Dibalik suksesnya ktt asean di labuan bajo, ada kekecewaan dari yang bertugas jaga keamanan disana. Udah capek tenaga, waktu, pikiran, dll tp semuanya gk dihargai. Cuma bisa ngeluh sesama temen yg tugas, gk bisa protes juga. Kasihan banget sumpah.	Neuroticism
ekasullistiani	Mau gimana juga kalo emg dasarnya udah egois, gk sadar diri ttep gak bakal pernah bisa berubah. Berujung jadi suka bohong.	Neuroticism
ekasullistiani	Hp lebih penting dari pada apapun😐	Neuroticism

4.3. *Preprocessing* Data dan *Split* Data

Setelah seluruh dataset dilabeli berdasarkan framework DISC dan framework OCEAN oleh pakar, tahapan selanjutnya merupakan melakukan

preprocessing data untuk mengelola data yang diperoleh agar dapat digunakan dalam meningkatkan kinerja model klasifikasi yang diusulkan. Proses *preprocessing* dilakukan melalui beberapa tahap yaitu

a. Case Folding

Pada proses ini semua tweet akan dilakukan beberapa aksi seperti mengubah semua tweet menjadi huruf kecil, menghapus username, menghapus tanda baca, dan menghapus *link* atau tautan. Gambar 4.2 merupakan hasil dari tahapan case folding pada dataset DISC.

Gambar 4.2. Hasil Tahapan Case Folding

b. Tokenization

Pada tahapan ini, kalimat pada setiap tweet akan diubah menjadi token-token. Hasil dari tahapan ini dapat dilihat pada Gambar 4.3

Gambar 4.3. Hasil Tahapan Tokenization

Pada Gambar 4.3 terlihat bahwa setelah melalui tahapan tokenization, kalimat kalimat yang ada pada satu tweet terurai menjadi beberapa token-token atau dipisah menjadi perkata.

c. Normalization

Pada proses ini terjadi penormalan sebuah kata atau kalimay dengan mengubah beberapa kalimat yang disingkat menjadi kata yang lebih baku. Pada tahapan ini, ditambahkan kamus yang berisi tentang kata-kata singkatan yang sering digunakan dan bentuk bakunya. Hasil dari tahapan ini dapat dilihat pada Gambar 4.4

Gambar 4.4. Hasil Tahapan Normalization

Pada Gambar 4.4 terlihat bahwa pada tahapan ini terdapat beberapa kata yang mengalami perubahan kata namun masih memiliki makna yang sama, seperti kata 'pengen' berubah menjadi 'ingin', kata 'bikin' menjadi 'buat; kata 'dengerin' menjadi 'mendengarkan' dan masih banyak lagi.

d. *Stopword removal*

Pada proses ini membuang kata-kata yang tidak penting atau yang tidak memiliki arti. Gambar 4.5 merupakan hasil dari tahapan *Stopword removal*.

Gambar 4.5. Hasil Tahapan *Stopword removal*

Pada Gambar 4.5 terlihat bahwa beberapa kata terlihat terhapus. Seperti pada kalimat pertama, tweet yang awalnya terdiri dari 10 kata, hanya tersisa 4 kata, begitupun dengan tweet yang lain.

e. Stemming

Stemming merupakan tahapan yang berfungsi untuk memberikan pemetaan pada sebuah kata-kata yang akan diubah menjadi kata dasar. Gambar 4.6 merupakan hasil dari tahapan *Stopword removal*.

Gambar 4.6. Hasil Tahapan *Stemming*

Pada Gambar 4.6 terlihat bahwa beberapa kata yang memiliki imbuhan mengalami pelepasan imbuhan. Seperti pada kalimat pertama, kata 'pesawatnya' menjadi 'pesawat'.

Setelah dilakukan *preprocessing*, tahapan selanjutnya adalah *split* data. Pada penelitian ini *split* data akan dilakukan dengan 4 kali percobaan yaitu dengan rasio 90:10, 80:10, 70:30, dan 60:40. Setelah data terbagi menjadi training dan testing, selanjutnya akan dilakukan pembobotan kata dengan *feature extraction*.

Pada penelitian ini, akan dilakukan tiga jenis *feature extraction* yaitu menggunakan TF-IDF, WF-IDF dan N-gram.

4.4. Clasification Model dan Evaluasi Model

Setelah dilakukan preprocessing, selanjutnya SVM akan membuat prediksi berdasarkan data yang sudah di preprocesisng dan sudah dilakukan feature extraction. Proses klasifikasi dilakukan pada setiap tweet. Setelah dilakukan proses klasifikasi menggunakan SVM, selanjutnya adalah dilakukan evaluasi dengan membandingkan hasil prediksi dengan label asli. Pada proses evaluasi ini, penelitian ini mengambil label yang paling sering muncul dari tweet yang sudah diklasifikasikan sebelumnya sebagai prediksi final untuk pengguna tersebut, sehingga hanya akan ada 52 data pada confussion matrix nantinya. Detail koding yang digunakan dapat dilihat pada Gambar 4.7.

```
# Mengambil label yang paling sering muncul di antara semua prediksi
def most_common_label(labels):
    return Counter(labels).most_common(1)[0][0]

# Membuat dataframe baru untuk menyimpan label yang paling sering muncul untuk setiap username
unique_users = test_df['ID'].unique()

# Inisialisasi list untuk menyimpan hasil prediksi dan label asli berdasarkan pengguna
all_y_true = []
all_y_pred = []

# Loop melalui setiap pengguna di data uji
for user in unique_users:
    # Filter data uji untuk pengguna tertentu
    user_tweets = test_df[test_df['ID'] == user]
    X_user_tfidf = tfidf_vectorizer.transform(user_tweets['tweet'])

    # Prediksi label untuk pengguna tersebut
    user_pred = svm_model.predict(X_user_tfidf)

    # Mengambil label yang paling sering muncul di antara semua prediksi
    predicted_label = most_common_label(user_pred)

    # Mengambil label asli dari baris pertama pengguna tersebut
    actual_label = most_common_label(user_tweets['Label'])

    # Tambahkan hasil prediksi dan label asli ke dalam list
    all_y_true.append(actual_label)
    all_y_pred.append(predicted_label)  # Menggunakan label yang paling sering muncul
```

Gambar 4.7. Detail Koding Proses Evaluasi

Pada Gambar 4.7 terlihat bahwa terdapat fungsi *most_common_label* digunakan untuk menentukan label yang paling sering muncul dalam daftar prediksi. Fungsi ini menghitung frekuensi setiap label dalam daftar dan mengembalikan label dengan frekuensi tertinggi. Ini penting ketika kita ingin merangkum beberapa prediksi menjadi satu keputusan akhir, seperti menentukan kategori keseluruhan untuk seorang pengguna berdasarkan semua tweet mereka.

Selanjutnya, kode ini mengidentifikasi pengguna unik dari dataset uji dan menyimpan mereka dalam variabel unique_users. Dua daftar kosong, all_y_true dan all_y_pred, diinisialisasi untuk menyimpan label asli dan label prediksi berdasarkan pengguna. Dengan mengisolasi pengguna, kode ini memungkinkan analisis lebih lanjut tentang bagaimana model bekerja pada tingkat pengguna, bukan hanya pada tingkat tweet individu.

Terakhir, untuk setiap pengguna, kode mengekstrak semua tweet mereka dari dataset uji, mengubah tweet tersebut menjadi vektor TF-IDF, dan membuat prediksi menggunakan model SVM. Label prediksi yang paling sering muncul di antara semua tweet pengguna tersebut kemudian disimpan sebagai prediksi akhir untuk pengguna itu. Hal yang sama dilakukan untuk label asli, dan hasilnya disimpan dalam all_y_true dan all_y_pred, yang dapat digunakan untuk evaluasi performa model secara keseluruhan. Hasilnya, label asli dan prediksi disimpan dalam dua daftar, yang memungkinkan analisis lebih lanjut tentang kinerja model pada tingkat pengguna, bukan hanya per tweet. Dengan pendekatan ini, model dapat memberikan prediksi yang dikonsolidasikan pada tingkat pengguna, bukan hanya pada tingkat tweet individu.

4.5. Skenario Percobaan

Pada penelitian ini terdapat beberapa skenario yang akan dilakukan. Hasil analisis dari beberapa skenario percobaan yang dilakukan akan digunakan sebagai acuan untuk mendapatkan akurasi terbaik. Detail skenario yang akan dilakukan pada penelitian ini dapat dilihat pada Tabel 4.7

Tabel 4.7. Skenario Penelitian

No	Skenario	Keterangan
1.	S1	SVM + *Preprocessing*
2.	S2	SVM + *Preprocessing* + *Split* Data
3.	S3	SVM + *Preprocessing* + *Split* Data + *Feature extraction*
4.	S4	SVM + *Preprocessing* + *Split* Data + *Feature extraction* + Kernel

Tabel 4.5 merupakan skenario yang digunakan pada penelitian ini untuk mencari akurasi terbaik. Penjelasan detail dari ketiga skenario pada Tabel 4.5 adalah sebagai berikut:

1. Skenario 1 (S1)

   Pada skenario pertama, dilakukan analisis terhadap pengaruh penggunaan *stopword removal* pada *preprocessing* data dalam melakukan klasifikasi teks menggunakan SVM. Parameter awal yang dipilih adalah menggunakan *feature extraction* TF-IDF dan pembagian data dengan rasio 80:20 untuk data training dan data testing. Pada skenario ini model SVM akan dibandingkan perfomanya ketika menggunakan *stopword removal* dan ketika tidak menggunakan *stopword removal* pada tahap *preprocessing*.

2. Skenario 2 (S2)

   Pada skenario kedua, dilakukan analisis terhadap pengaruh pemilihan pembagian data training dan testing dalam melakukan klasifikasi teks menggunakan SVM. Pada skenario 2 ini, tahapan *preprocessing* yang digunakan merupakan tahapan *preprocessing* yang mendapat akurasi tertinggi pada skenario sebelumnya, sedangkan *feature extraction* yang digunakan merupakan TF-IDF. Terdapat empat percobaan yang dilakukan pada pembagian data ini, yaitu pembagian data 90:10, 80:20, 70:30, dan 60:40.

3. Skenario 3 (S3)

   Pada skenario ketiga, dilakukan analisis terhadap pengaruh pemilihan *feature extraction* dalam melakukan klasifikasi teks menggunakan SVM. Pada skenario 3 ini, tahapan *preprocessing* dan pembagian dataset yang digunakan merupakan tahapan yang mendapat hasil akurasi tertinggi dari skenario sebelumnya. Pada skenario ini, *feature extraction* yang digunakan terdiri dari 3 percobaan yaitu menggunakan TF-IDF, WF-IDF, dan N-gram.

4. Skenario 4 (S4)

   Pada skenario keempat, dilakukan analisis terhadap pengaruh pemilihan kernel SVM dalam melakukan klasifikasi teks. Pada tahapan ini, tahapan *preprocessing*, pembagian dataset dan *feature extraction* yang digunakan merupakan hasil terbaik dari skenario sebelumnya.

Kernel yang akan di uji cobakan pada skenario ini adalah linear, rbf, sigmoid, dan polynomial.

4.6. Hasil dan Evaluasi

Proses klasifikasi penyakit pada daun teh dilakukan sesuai dengan skenario yang telah ditetapkan sebelumnya. Hasil percobaan yang dilakukan akan dijelaskan secara rinci dan dilakukan analisis terhadap pengaruh penggunaan *stopword removal* pada *preprocessing* data, pemilihan *split* data, pemilihan *feature extraction*, dan pemilihan kernel SVM.

4.4.1 Hasil dan Evaluasi Skenario 1

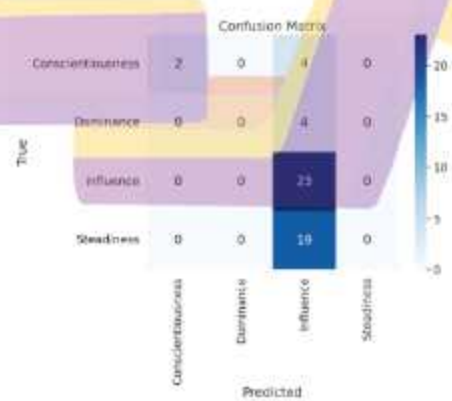
Pada skenario pertama, dilakukan analisis terhadap pengaruh penggunaan *preprocessing* data dalam melakukan klasifikasi teks menggunakan SVM. *Preprocessing* data yang dilakukan terdiri dari lima tahapan, yaitu case folding, tokenization, normalization, *stopword removal*, dan stemming. Namun, pada skenario pertama ini, terdapat dua hal yang akan dilakukan, yaitu menggunakan kelima tahapan *preprocessing* data tersebut dan hanya akan menggunakan 4 tahapan dengan menghilangkan tahapan *stopword removal*

Parameter awal yang dipilih adalah menggunakan *feature extraction* TF-IDF dan pembagian data 80:20. Pada skenario ini model SVM akan dijalankan pada dataset DISC dan OCEAN. Untuk hasil akhir yang akan ditampilkan merupakan akurasi SVM untuk setiap pengguna dari masing-masing dataset yang digunakan. Tabel 4.8 merupakan hasil perbandingan akurasi dari penggunaan *stopword removal* dan tanpa penggunaan *stopword removal* pada proses *preprocessing* data.

Tabel 4.8. Hasil Skenario 1

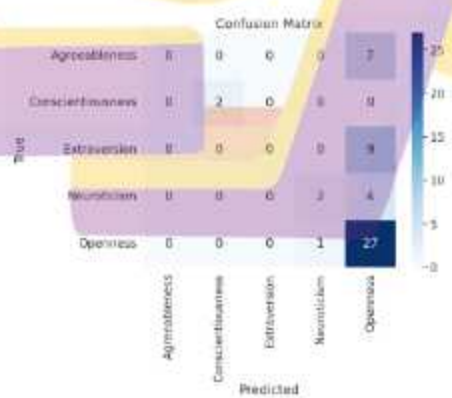
Dataset	DISC				OCEAN			
	Akurasi	Presisi	Recall	F1Score	Akurasi	Presisi	Recall	F1Score
Dengan Stopword	0,44	0,11	0,25	0,15	0,54	0,16	0,22	0,18
Tanpa Stopword	**0,48**	**0,36**	**0,33**	**0,28**	**0,60**	**0,45**	**0,46**	**0,43**

Berdasarkan pada Tabel 4.8, terlihat bahwa hasil optimal untuk kedua dataset adalah tanpa menggunakan *stopword removal* pada tahapan *preprocessing*. Pada dataset DISC, diperoleh nilai akurasi tertinggi sebesar 48%, presisi 36%, recall 33% dan F1Score 28%, sedangkan pada dataset OCEAN diperoleh hasil akurasi tertinggi sebesar 60%, presisi 45%, recall 46%, dan F1Score 43%. Karena penggunaan tanpa *stopword removal* ini memperoleh akurasi tertinggi untuk kedua dataset, sehingga pada skenario selanjutnya kedua dataset tidak akan menggunakan *stopword removal* pada proses *preprocessing* data. Confusion matrix untuk dataset DISC dapat dilihat pada Gambar 4.8.



Gambar 4.8. Confusion Matrix Skenario 1 Dataset DISC

Pada Gambar 4.8 terlihat bahwa terdapat 25 user yang terklasifikasi dengan benar dan 27 user yang diklasifikasikan dengan salah. Pada kelas Conscientiousmess terdapat dua user yang diklasifikasikan dengan benar dan terdapat 4 user yang salah diklasifikasikan sebagai kelas *influence*. Pada kelas *Dominance*, tidak terdapat user yang diklasifikasikan dengan benar karena ke empat user yang masuk ke dalam kelas *dominance* salah diklasifikasikan sebagai kelas *influence*. Pada kelas *Influence*, semua user sejumlah 23 user diklasifikasikan dengan benar sebagai kelas *influence*. Terakhir, pada kelas *steadiness*, semua user sejumlah 19 user salah diklasifikasikan sebagai kelas *influence*. Hasil akhir pada confusion matrix menjadi berbeda dengan yang ada pada tabel 4.3 karena adanya proses pembagian data. Setelah terjadi proses split data dan klasifikasi, kelas Conscientiousmess pada skenario ini terdiri dari 6 user, kelas dominance terdiri dari 4 user, influence 23 user, dan steadiness 19 user. Selanjutnya confusion matrix untuk dataset OCEAN dapat dilihat pada Gambar 4.9.



Gambar 4.9. Confusion Matrix Skenario 1 Dataset OCEAN

Pada Gambar 4.9 terlihat bahwa terdapat 31 user yang terklasifikasi dengan benar dan 21 user yang diklasifikasikan dengan salah. Pada kelas *Agreeableness,* semua user sejumlah 7 user salah diklasifikasikan sebagai kelas *Openness*. Pada kelas *Conscientiousness* semua user sejumlah 2 user terklasifikasikan dengan benar. Pada kelas *Extraversion* smua user sejumlah 9 user salah diklasifikasikan sebagai kelas *Openness*. Pada kelas *Neuroticism* terdapat 2 user yang terklasifikasikan dengan benar dan 4 user lainnya salah diklasifikasikan sebagai kelas *Openness*. Terakhir pada kelas *openness*, terdapat 27 akun yang diklasifikasikan dengan benar dan 1 akun yang salah diklasifikasikan sebagai kelas *Neuroticism*. Hasil akhir pada confusion matrix menjadi berbeda dengan yang ada pada tabel 4.4 karena adanya proses pembagian data yang berbeda dari sebelumnya. Setelah terjadi proses split data dan klasifikasi terjadi perubahan jumlah data pada tiap tiap kelas, kelas *Agreeableness* terdiri dari 7 user, kelas Conscientiousmess terdiri dari 2 user (hanya kelas ini yang tidak mengalami perubahan), kelas Extraversion terdiri dari 9 user, kelas neuroticism terdiri dari 6 user, dan terakhir kelas *Openness* terdiri dari 28 user.

Tahapan preprocessing tanpa menggunakan stopword removal terbukti menghasilkan akurasi yang lebih tinggi dibanding ketika menggunakan stopword removal. Hal ini menunjukkan bahwa pada dataset yang digunakan, baik DISC maupun OCEAN, mempertahankan stopword bisa lebih bermanfaat daripada menghapusnya. Sebagai contoh, dalam klasifikasi sentimen, kata-kata seperti "tidak", "di", "aku" atau "wkwkwk" yang sering kali dianggap sebagai stopword justru dapat membawa makna penting dalam konteks kalimat. Menghapus kata-kata

ini bisa menghilangkan informasi yang krusial untuk model dalam memahami sentimen yang sebenarnya. Pada kalimat "Aku punya kucing putih di rumah dan emang bego sih wkwkwkwk", akan di labeli awal oleh psikolog sebagai kelas Influence karena kalimatnya menggunakan gaya yang santai dan humoris. ditandai dengan penggunaan "wkwkwkwk." Ini menunjukkan sifat ekstrovert, ramah, dan ekspresif yang biasa diasosiasikan dengan tipe kepribadian Influence. Sedangkan ketika *stopword removal* digunakan, kalimatnya akan menjadi "punya kucing putih rumah emang bego". Kalimat tersebut ditandai sebagai dominance karena kalimatnya tegas diucap secara langsung, singkat, dan to the point, tanpa banyak detail atau ekspresi tambahan. Hal itulah yang menyebabkan pada beberapa kelas, kalimat salah diklasifikasikan ketika menghilangkan stopword (*stopword removal*). Oleh karena itu, dalam beberapa kasus, mempertahankan stopword atau kata kata umum dapat membantu model untuk lebih akurat mengidentifikasi pola, yang pada akhirnya dapat meningkatkan nilai akurasi.

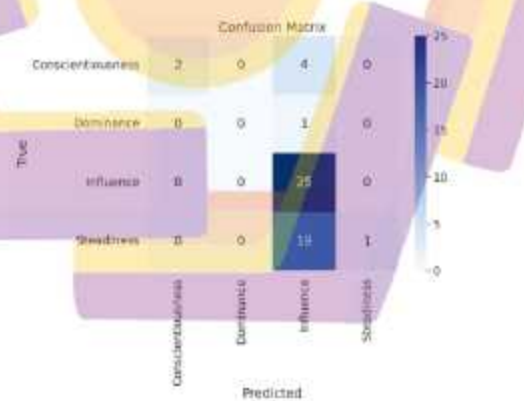
4.4.2 Hasil dan Evaluasi Skenario 2

Pada skenario kedua, dilakukan analisis terhadap pengaruh pemilihan pembagian data training dan testing dalam melakukan klasifikasi teks menggunakan SVM. Pada tahapan ini, tahapan *preprocessing* yang digunakan merupakan tahapan *preprocessing* yang mendapat akurasi tertinggi yaitu tanpa menggunakan *stopword removal*, sedangkan *feature extraction* yang digunakan merupakan TF-IDF. Terdapat empat percobaan yang dilakukan pada pembagian data ini, yaitu pembagian data 90:10, 80:20, 70:30, dan 60:40. Tabel 4.7 merupakan hasil perbandingan akurasi pada skenario pembagian data.

Tabel 4.9. Hasil Skenario 2

Skenario	DISC				OCEAN			
	Akurasi	Presisi	Recall	F1Score	Akurasi	Presisi	Recall	F1Score
90 10	0,46	0,26	0,28	0,21	0,54	0,73	0,48	0,44
80 20	0,48	0,36	0,33	0,28	0,6	0,45	0,46	0,43
70 30	**0,54**	**0,62**	**0,34**	**0,32**	0,58	0,71	0,45	0,42
60 40	0,50	0,27	0,35	0,30	**0,65**	**0,92**	**0,49**	**0,50**

Berdasarkan pada Tabel 4.9 terlihat bahwa kedua dataset mencapai hasil yang optimal pada pembagian data yang berbeda. Pada dataset DISC, hasil tertinggi diperoleh ketika pembagian data dilakukan dengan perbandingan 70 untuk data training dan 30 untuk data testing dengan akurasi 54%, presisi 62%, recall 34%, dan F1 Score 32%. Sedangkan pada dataset OCEAN, diperoleh hasil yang optimal ketika pembagian data dilakukan dengan perbandingan 60 untuk data training dan 40 untuk data testing dengan akurasi 65%, presisi 92%, recall 49%, dan F1Score 50%. Confusion matrix untuk dataset DISC dapat dilihat pada Gambar 4.10.



Gambar 4.10. Confusion Matrix Skenario 2 Dataset DISC

Pada Gambar 4.10 terlihat bahwa terdapat 28 user yang terklasifikasi dengan benar dan 24 user yang diklasifikasikan dengan salah. Pada kelas

Conscientiousmess terdapat 2 user yang diklasifikasikan dengan benar dan 4 user yang salah diklasifikasikan sebagai kelas *influence*. Pada kelas *Dominance*, tidak terdapat user yang diklasifikasikan dengan benar karena 1 user yang masuk ke dalam kelas *dominance* salah diklasifikasikan sebagai kelas *influence*. Pada kelas *Influence*, semua user sejumlah 25 user diklasifikasikan dengan benar sebagai kelas *influence*. Terakhir, pada kelas *steadiness*, terdapat 1 user yang benar diklasifikasikan sebagai kelas *steadiness* dan 19 user salah diklasifikasikan sebagai kelas *influence*. Hasil akhir pada confusion matrix menjadi berbeda dengan yang ada pada tabel 4.3 karena adanya proses pembagian data yang berbeda dari sebelumnya. Setelah terjadi proses split data dan klasifikasi, kelas Conscientiousmess pada skenario ini terdiri dari 6 user, kelas dominance terdiri dari 1 user, influence 25 user, dan steadiness 20 user.

Pada skenario 2 ini, dataset DISC memperoleh akurasi tertinggi ketika menggunakan pembagian data dengan rasio 70:30. Hal ini dapat disebabkan karena adanya lebih banyak data yang dialokasikan untuk pengujian, sehingga model memiliki lebih banyak data untuk diuji, yang dapat membantu mengevaluasi kinerjanya dengan lebih baik. Terlebih pada proses klasifikasi kepribadian menggunakan dataset DISC ini, akurasi dihitung berdasarkan akumulasi perhitungan setiap satu user yang terdiri dari beberapa tweet didalamya. Sehingga data uji yang lebih banyak mampu memberikan hasil akumulasi yang lebih tepat. Selanjutnya confusion matrix untuk dataset OCEAN dapat dilihat pada Gambar 4.11.

Gambar 4.11. Confusion Matrix Skenario 2 Dataset OCEAN

Pada Gambar 4.11 terlihat bahwa terdapat 34 user yang terklasifikasi dengan benar dan 18 user yang diklasifikasikan dengan salah. Pada kelas *Agreeableness,* , 1 user diklasifikasikan benar sebagai kelas *Agreeableness,* dan 6 user salah diklasifikasikan sebagai kelas *Openness*. Pada kelas *Conscientiousness* semua user sejumlah 2 user terklasifikasikan dengan benar. Pada kelas *Extraversion* 1 user diklasifikasikan dengan benar sebagai kelas *extraversion* dan 5 user salah diklasifikasikan sebagai kelas *Openness*. Pada kelas *Neuroticism* terdapat 1 user yang terklasifikasikan dengan benar dan 7 user lainnya salah diklasifikasikan sebagai kelas *Openness*. Terakhir pada kelas *openness*, terdapat 29 akun yang diklasifikasikan dengan benar sebagai kelas *openness*. Setelah terjadi proses split data yang berbeda, terjadi perbedaan kembali pada beberapa kelas. Pada kelas *Agreeableness* terdiri dari 7 user, kelas Conscientiousmess terdiri dari 2 user, kelas Extraversion terdiri dari 6 user, kelas neuroticism terdiri dari 8 user, dan terakhir kelas *Openness* terdiri dari 28 user.

Alokasi data mempengaruhi kinerja model dengan menentukan seberapa baik model dapat belajar dan menggeneralisasi. Ketika sebagian besar data digunakan untuk pelatihan, model dapat mempelajari pola dan fitur dari data dengan lebih mendalam, mengurangi risiko underfitting. Namun, alokasi yang sangat besar untuk pelatihan dan sangat sedikit untuk pengujian dapat menyebabkan overfitting, di mana model terlalu spesifik pada data latih dan kurang efektif pada data baru. Sebaliknya, alokasi yang lebih seimbang antara data pelatihan dan data uji memungkinkan model untuk belajar secara efektif sambil memastikan bahwa data uji cukup representatif untuk mengevaluasi kinerja model secara akurat. Dengan keseimbangan yang baik, evaluasi model akan lebih stabil dan hasilnya lebih dapat diandalkan dalam mencerminkan performa model pada data yang belum pernah dilihat sebelumnya.

Pada skenario 2 ini, dataset DISC memperoleh hasil tertinggi ketika menggunakan alokasi data 70:30, sedangkan pada OCEAN diperoleh hasil tertinggi ketika menggunakan alokasi data 60:40. Hasil akurasi yang lebih tinggi pada rasio 70:30 dan 60:40 dibandingkan dengan 90:10 dan 80:20 disebabkan oleh cara klasifikasi yang dilakukan dalam menangani prediksi per pengguna yang dijelaskan pada Gambar 4.7. Pada rasio 70:30 dan 60:40, model mendapatkan lebih banyak data latih, yang memungkinkan model untuk belajar lebih baik dari variasi dalam data. Dengan data latih yang lebih banyak, model bisa lebih efektif dalam memprediksi label untuk setiap tweet dan konsolidasi hasil prediksi per pengguna menjadi lebih akurat. Data uji yang lebih besar pada rasio ini juga memberikan gambaran yang lebih representatif tentang kinerja model, menghasilkan evaluasi

yang lebih stabil. Sebaliknya, rasio 90:10 dan 80:20 dengan data uji yang lebih kecil mungkin memberikan hasil yang kurang konsisten dan tidak mencerminkan performa model secara akurat karena ukuran data uji yang tidak memadai.

4.4.3 Hasil dan Evaluasi Skenario 3

Pada skenario ketiga, dilakukan analisis terhadap pengaruh pemilihan *feature extraction* dalam melakukan klasifikasi teks menggunakan SVM. Pada tahapan ini, tahapan *preprocessing* yang digunakan merupakan tahapan *preprocessing* yang mendapat akurasi tertinggi. Pembagian dataset yang digunakan juga merupakan pembagian dataset yang menghasilkan akurasi tertinggi. Karena terdapat perbedaan hasil terbaik pada pembagian data, selanjutnya untuk dataset DISC akan menggunakan pembagian data 70:30 sedangakan dataset OCEAN akan menggunakan pembagian data 60:40.

Pada skenario ini, *feature extraction* yang digunakan terdiri dari 3 percobaan yang dilakukan yaitu menggunakan TF-IDF, WF-IDF, dan N-gram. Tabel 4.10 merupakan hasil perbandingan akurasi pada skenario pemilihan *feature extraction*.

Tabel 4.10. Hasil Skenario 3

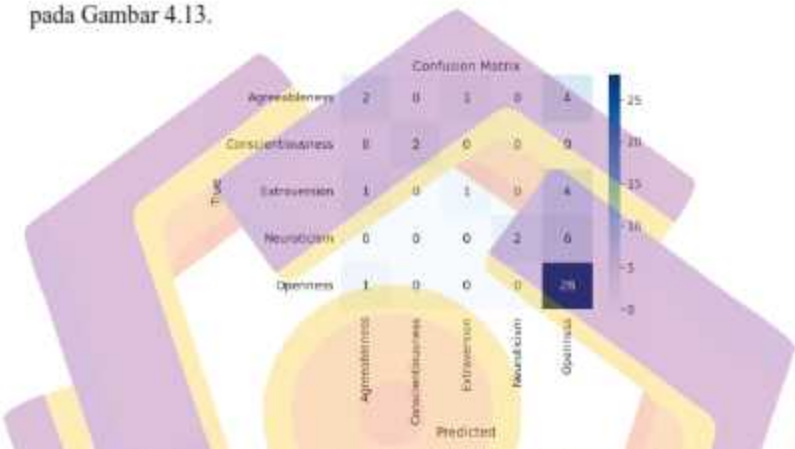
Skenario	DISC (70:30)				OCEAN (60:40)			
	Akurasi	Presisi	Recall	F1Score	Akurasi	Presisi	Recall	F1Score
TFIDF	0,54	0,62	0,34	0,32	0,65	0,92	0,49	0,5
WFIDF	0,52	0,37	0,36	0,33	0,4	0,65	0,48	0,45
N-gram (1,1)	0,52	0,37	0,33	0,292	**0,67**	**0,73**	**0,53**	**0,56**
N-gram (1,2)	**0,56**	**0,63**	**0,36**	**0,34**	0,63	0,92	0,39	0,43
N-gram (1,3)	0,5	0,37	0,26	0,19	0,62	718	0,29	0,3
N-gram (2,2)	0,48	0,12	0,25	0,16	0,58	0,31	0,23	0,2
N-gram (2,3)	0,48	0,12	0,25	0,16	0,56	0,11	0,2	0,14
N-gram (3,3)	0,48	0,12	0,25	0,16	0,56	0,11	0,2	0,14

Pada Tabel 4.10 terlihat bahwa hasil kedua dataset memperoleh hasil yang optimal pada penggunaan n-gram. Namun, perlu dilihat kembali bahwa terdapat perbedaan nilai n pada kedua dataset. Pada dataset DISC, diperoleh hasil yang maksimal ketika menggunakan kombinasi unigram dan bigram atau biasa disebut dengan n-gram (1,2), yaitu dengan akurasi 56%, presisi 63%, recall 36%, dan F1Score 34%. Sedangkan pada dataset OCEAN diperoleh hasil maksimal ketika hanya menggunakan unigram atau n-gram (1,1), yaitu dengan akurasi 67%, presisi 73%, recall 53%, dan F1Score 56%. Confusion matrix untuk dataset DISC dapat dilihat pada Gambar 4.12.

Gambar 4.12. Confusion Matrix Skenario 3 Dataset DISC

Pada Gambar 4.12 terlihat bahwa terdapat 28 user yang terklasifikasi dengan benar dan 24 user yang diklasifikasikan dengan salah. Pada kelas *Conscientiousness* terdapat 2 user yang diklasifikasikan dengan benar dan 4 user yang salah diklasifikasikan sebagai kelas *influence*. Pada kelas *Dominance*, tidak terdapat user yang diklasifikasikan dengan benar karena 1 user yang masuk ke dalam kelas *dominance* salah diklasifikasikan sebagai kelas *influence*. Pada kelas

Influence, semua user sejumlah 25 diklasifikasikan dengan benar sebagai kelas *influence*. Terakhir, pada kelas *steadiness*, terdapat 2 user yang benar diklasifikasikan sebagai kelas *steadiness* dan 18 user salah diklasifikasikan sebagai kelas *influence*. Selanjutnya confusion matrix untuk dataset OCEAN dapat dilihat pada Gambar 4.13.



Gambar 4.13. Confusion Matrix Skenario 3 Dataset OCEAN

Pada Gambar 4.13 terlihat bahwa terdapat 35 user yang terklasifikasi dengan benar dan 17 user yang diklasifikasikan dengan salah. Pada kelas *Agreeableness*, 2 user diklasifikasikan benar sebagai kelas *Agreeableness*, dan terdapat 2 kesalahan dimana 1 user salah diklasifikasikan sebagai kelas *Extraversion* dan 4 user salah diklasifikasikan sebagai kelas *Openness*. Pada kelas *Conscientiousnesss* emua user sejumlah 2 user terklasifikasikan dengan benar. Pada kelas *Extraversion* 1 user diklasifikasikan dengan benar sebagai kelas *extraversion* dan terdapat 2 kesalahan dimana 1 user salah diklasifikasikan sebagai kelas *Agreeableness*, dan 4 user salah diklasifikasikan sebagai kelas *Openness*. Pada kelas *Neuroticism* terdapat 2 user yang terklasifikasikan dengan benar dan 6 user

lainnya salah diklasifikasikan sebagai kelas *Openness*. Terakhir pada kelas *openness*, terdapat 29 akun yang diklasifikasikan dengan benar sebagai kelas *openness*

Ekstraksi fitur menggunakan TF-IDF, WF-IDF, dan n-gram secara efektif meningkatkan kemampuan model machine learning dalam analisis teks. TF-IDF dan WF-IDF memberikan bobot pada kata-kata berdasarkan frekuensi kemunculannya dalam dokumen dan seberapa jarang kata tersebut muncul di seluruh koleksi dokumen, sehingga membantu model untuk fokus pada kata-kata yang lebih relevan dan signifikan. Di sisi lain, n-gram menangkap konteks lokal dengan mempertimbangkan urutan kata, memungkinkan model untuk memahami frasa atau pola kata yang sering muncul bersama. Kombinasi dari ketiga metode ini memberikan representasi teks yang lebih informatif dan kontekstual, meningkatkan akurasi dan efektivitas model dalam analisis atau klasifikasi teks.

Pada skenario ini, kedua dataset memiliki akurasi tertinggi ketika menggunakan n-gram dengan nilai n = 1. Hal ini menunjukkan bahwa representasi teks menggunakan kata tunggal sudah cukup untuk model dalam konteks ini, karena kata-kata individual memberikan informasi yang relevan dan memadai untuk tugas klasifikasi atau analisis. Penggunaan unigrams memungkinkan model untuk fokus pada kata-kata yang paling signifikan tanpa terpengaruh oleh kompleksitas tambahan dari n-gram dengan nilai n yang lebih besar. Dalam kasus ini, unigrams sudah efektif dalam membedakan kelas atau kategori dalam data, menunjukkan bahwa fitur dari kata tunggal sudah mencakup informasi yang diperlukan untuk membuat prediksi yang akurat.

4.4.4 Hasil dan Evaluasi Skenario 4

Pada skenario keempat, dilakukan analisis terhadap pengaruh pemilihan kernel SVM dalam melakukan klasifikasi teks. Pada tahapan ini, tahapan *preprocessing* yang digunakan merupakan tahapan *preprocessing* yang mendapat akurasi tertinggi. Pembagian dataset yang digunakan juga merupakan pembagian dataset yang menghasilkan akurasi tertinggi. *Feature extraction* yang digunakan juga merupakan *feature extraction* yang menghasilkan akurasi tertinggi dari skenario sebelumnya. Karena terdapat perbedaan nilai n pada n*feature extraction* n-gram, pada skenario 4 ini, dataset DISC akan menggunakan n-gram (1,2) atau kombinasi antara unigram dan bigram sedangkan dataset OCEAN akan menggunakan n-gram (1,1) atau hanya menghasilkan unigram (satu item).

Kernel yang akan di uji cobakan pada skenario ini adalah linear, rbf, sigmoid, dan polynomial. Pada kernel linear dan sigmoid, proses klasifikasi bisa dilakukan secara langsung tanpa menambahkan parameter lain. Sedangkan pada kernel rbf dan polynomial, dibutuhkan parameter nilai C dan gamma. Pada penelitian ini, dilakukan pengadopsian teknik GridSearchCV pada kernel rbf dan kernel polynomial untuk menguji setiap kombinasi dari parameter nilai C dan gamma untuk menentukan kombinasi yang menghasilkan kinerja terbaik pada model SVM sehingga kombinasi nilai C dan gamma untuk setiap dataset pada kernel rbf dan kernel polinomial akan menggunakan kombinasi yang terbaik menurut hasil GridSearchCV. Tabel 4.11 merupakan hasil perbandingan akurasi dari pemilihan kernel

Tabel 4.11. Hasil Skenario 4

Skenario	DISC (70:30) + N-gram (1,2)				OCEAN (60:40) + N-gram (1,1)			
	Akurasi	Presisi	Recall	F1Score	Akurasi	Presisi	Recall	F1Score
Linear	**0,56**	**0,63**	**0,36**	**0,34**	**0,67**	**0,73**	**0,53**	**0,56**
RBF	0,54	0,38	0,29	0,25	0,63	0,92	0,39	0,44
Polynomial	0,48	0,12	0,25	0,16	0,60	0,52	0,26	0,25
Sigmoid	0,54	0,13	0,25	0,17	0,56	0,11	0,20	0,14

Berdasarkan pada Tabel 4.9 terlihat bahwa kedua dataset memiliki hasil yang optimal yang sama yaitu pada penggunaan kernel linear, dimana pada DISC diperoleh akurasi 56%, presisi 63%, recall 36%, dan F1Score 34%. Sedangkan pada dataset OCEAN diperoleh akurasi 67%, presisi 73%, recall 53%, dan F1Score 56%. Karena hasil yang diperoleh sama dengan skenario sebelumnya yaitu tetap menggunakan linear, confusion matrix untuk dataset DISC dapat dilihat pada Gambar 4.10 dan confusion matrix skenario 4 untuk dataset OCEAN dapat dilihat pada Gambar 4.11.

Kernel dalam Support Vector Machine (SVM) mempengaruhi bagaimana data dipetakan ke ruang fitur yang lebih tinggi untuk memudahkan pemisahan antar kelas. Kernel Linear cocok untuk data yang dapat dipisahkan secara linier, memberikan efisiensi dan performa baik pada kasus linear. Kernel Polinomial menangani data non-linier dengan memetakan data ke ruang fitur yang lebih tinggi berdasarkan derajat polinomial, memungkinkan model untuk menangkap pola yang lebih kompleks namun memerlukan penyesuaian parameter yang hati-hati. Kernel Radial Basis Function (RBF) sangat fleksibel dan dapat menangani pola data yang sangat kompleks dan non-linier, dengan parameter gamma yang menentukan jangkauan pengaruh data, tetapi memerlukan penyesuaian yang cermat untuk menghindari overfitting. Kernel Sigmoid, mirip dengan fungsi aktivasi dalam

jaringan saraf, juga tersedia tetapi kurang umum digunakan karena sering kalah saing dengan kernel linear dan RBF dalam performa. Pemilihan kernel yang tepat adalah kunci untuk meningkatkan akurasi dan efisiensi model SVM.

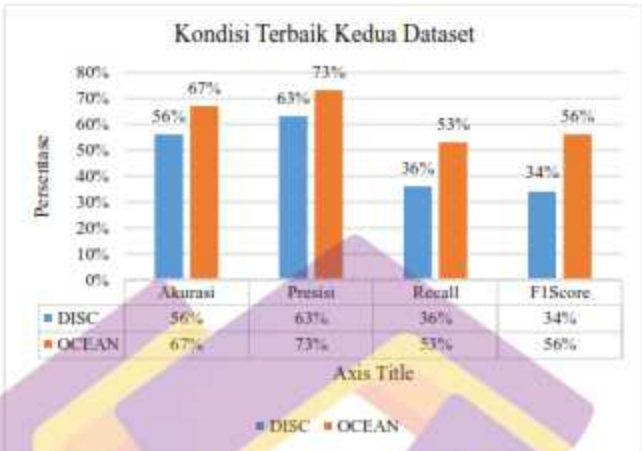
Hasil penelitian pada skenario 4 menunjukkan bahwa kernel linear menghasilkan akurasi tertinggi, yang mengindikasikan bahwa data yang dianalisis mungkin memiliki hubungan yang dapat dipisahkan secara linier dengan baik. Kernel linear menjadi lebih efektif ketika data dapat dipisahkan dengan garis lurus dalam ruang fitur yang ada, menunjukkan bahwa fitur-fitur dalam dataset sudah cukup representatif tanpa memerlukan pemetaan ke ruang fitur yang lebih tinggi. Kemungkinan lainnya adalah bahwa kompleksitas tambahan dari kernel polinomial atau RBF tidak memberikan manfaat signifikan dalam kasus ini. Kernel linear memungkinkan model SVM beroperasi dengan efisien tanpa risiko overfitting atau underfitting, terutama ketika data tidak memiliki pola yang sangat non-linier. Hasil ini menunjukkan bahwa kernel linear adalah opsi optimal untuk dataset ini, memberikan keseimbangan terbaik antara performa dan efisiensi dalam klasifikasi.

4.7. Kondisi Terbaik

Setelah dilakukan semua uji coba pada skenario, perubahan yang dihasilkan dari satu skenario dengan skenario yang lain ternyata tidak menghasilkan perbedaan hasil yang signifikan. Beberapa kelas memiliki akurasi yang rendah. Hal ini terjadi karena kondisi dataset yang memang tidak seimbang baik dari jumlah tweet maupun jumlah akun. Dapat dilihat pada Tabel 4.1, pada framework DISC khusunya kelas 'Influence' memang memiliki jumlah terbanyak dari segi tweet maupun jumlah akun, sehingga kelas tersebut sering diklasifikasikan dengan benar

dibanding kelas lain. Sedangkan pada kelas 'dominance' karena hanya ada satu akun yang terlabeli sebagai dominance dan hanya 318 tweet, membuat kelas ini sering sekali diklasifikasikan dengan salah. Hal tersebut juga terjadi pada Tabel 4.2 Terlihat pada framework OCEAN, kelas 'Openness' memiliki jumlah tweet dan akun terbanyak, sehingga kelas ini sering diklasifikasikan dengan benar dari setiap skenario, sedangkan kelas 'Conscientiousness' memiliki jumlah tweet dan akun paling sedikit sehingga sering salah diklasifikasikan sebagai kelas lain

Namun, pada kedua framework algoritma yang digunakan tetap bisa melakukan proses klasifikasi. Pada farmework DISC diperoleh akurasi 56%, presisi 63%, recall 36%, dan F1Score 34% dengan *preprocessing* data tanpa *stopword removal*, pembagian dataset dengan rasio 70:30, *feature extraction* n-gram (1,2), dan kernel SVM linear. Sedangkan pada dataset OCEAN diperoleh akurasi 67%, presisi 73%, recall 53%, dan F1Score 56% dengan *preprocessing* data tanpa *stopword removal*, pembagian dataset dengan rasio 60:40, *feature extraction* n-gram (1,1), dan kernel SVM linear. Gambar 4,14 merupakan perbandingan pada kondisi terbaik dari kedua dataset



Gambar 4.14. Perbandingan Kondisi Terbaik Kedua Dataset

Pada gambar 4.14, terlihat bahwa pada kedua dataset memiliki nilai presisi yang tinggi untuk kedua dataset yaitu 63% untuk kelas DISC dan 73% untuk kelas OCEAN. Hal ini dikarenakan dataset yang tidak seimbang. Presisi tinggi pada dataset yang tidak seimbang dapat terjadi karena model cenderung menghindari kesalahan prediksi positif palsu atau *false positives* dengan sangat efektif, terutama ketika ada kelas yang memiliki jauh lebih sedikit contoh dibandingkan kelas yang lain. Dalam situasi ini, model mungkin lebih sering memprediksi kelas mayoritas (seperti kelas *influnce* pada DISC dan *openness* pada OCEAN) untuk mengurangi kemungkinan membuat kesalahan, sehingga kelas lain yang memiliki nilai kemunculan lebih sedikit akan diklasifikasikan sebagai kelas mayoritas. Akibatnya, dari semua prediksi positif yang dibuat, sebagian besar benar, yang menghasilkan nilai presisi yang tinggi. Namun, ini sering datang dengan hasil recall yang rendah karena banyak contoh dari kelas yang jarang muncul (sedikit) mungkin tidak terdeteksi oleh model.

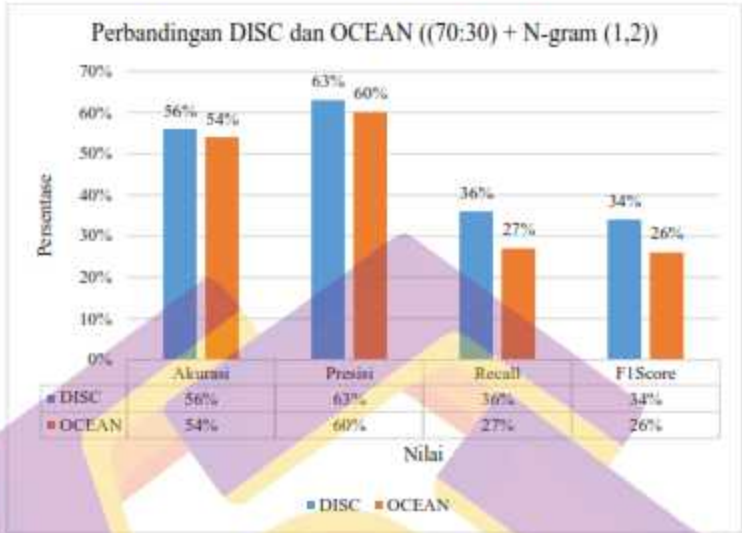
Pada Gambar 4.14, juga terlihat bahwa dataset OCEAN menghasilkan nilai akurasi yang lebih tinggi dibanding dengan dataset DISC. Namun, nilai akurasi yang dihasilkan memiliki perbedaan pada parameter rasio pembagian dataset dan kombinasi nilai n pada n-gram. Sehingga, perlu dilakukan pengujian kembali dengan menukar parameter yang digunakan pada dataset OCEAN dan dataset DISC untuk melihat sejauh mana performa SVM pada kedua dataset yang digunakan.

Pada pengujian ini, dataset DISC akan menggunakan rasio pembagian data 60:40 dan *feature extraction* n-gram (1,1), sedangkan dataset OCEAN akan menggunakan rasio pembagian data 70:30 dan *feature extraction* n-gram (1,2). Tabel 4.12 merupakan hasil perbandingan akurasi dari skenario tambahan dengan melakukan pertukaran parameter dari setiap dataset.

Tabel 4.12. Hasil Skenario Tambahan

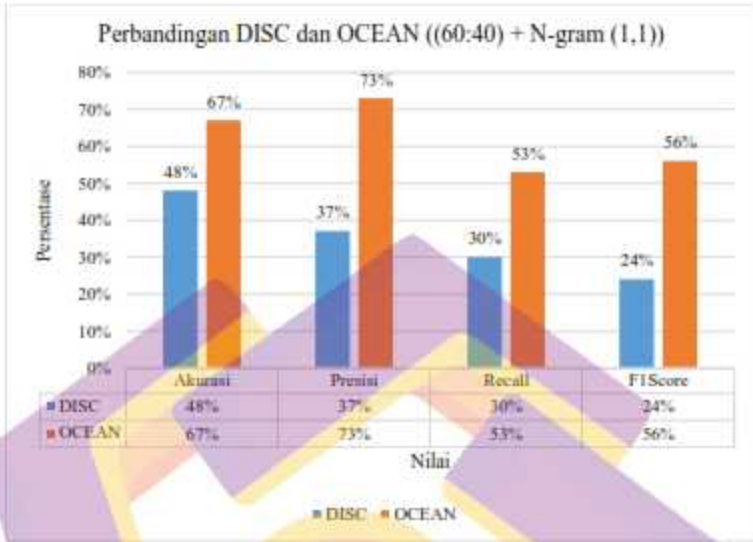
Skenario	Akurasi	Presisi	Recall	F1Score
DISC (70:30) + N-gram (1,2)	0,56	0,63	0,36	0,34
OCEAN (60:40) + N-gram (1,1)	0,67	0,73	0,53	0,56
DISC (60:40) + N-gram (1,1)	0,48	0,37	0,30	0,24
OCEAN (70:30) + N-gram (1,2)	0,54	0,60	0,27	0,26

Berdasarkan dari Tabel 4.12, terlihat bahwa ketika menggunakan rasio pembagian dataset 70:30 dan *feature extraction* n-gram (1,2), dataset OCEAN memiliki akurasi yang lebih rendah sebesar 0,02 dibanding dengan dataset DISC. Perbandingan akurasi dari dataset DISC dan dataset OCEAN menggunakan rasio pembagian dataset 70:30 dan *feature extraction* n-gram (1,2) dapat dilihat lebih detail pada Gambar 4.15.



Gambar 4.15. Perbadingan Dataset DISC dan Dataset OCEAN pada rasio 70:30 dan *feature extraction* n-gram (1,2)

Berdasarkan Gambar 4.15, Dataset DISC mendapatkan akurasi yang lebih tinggi dibanding dataset OCEAN dengan selisih 2% dimana DISC mendapat akurasi 56%, presisi 63%, recall 36%, dan F1Score 34%. Sedangkan dataset OCEAN mendapat akurasi 54%, presisi 60%, recall 27%, dan F1Score 26%. Selanjutnya, perbandingan akurasi dari dataset DISC dan dataset OCEAN menggunakan rasio pembagian dataset 60:40 dan *feature extraction* n-gram (1,1) dapat dilihat lebih detail pada Gambar 4.16.



Gambar 4.16. Perbadingan Dataset DISC dan Dataset OCEAN pada rasio 60:40 dan *feature extraction* n-gram (1,1).

Berdasarkan Gambar 4.16, Dataset OCEAN mendapatkan akurasi yang lebih tinggi dibanding dataset DISC dengan selisih yang cukup signifikan yaitu 19% dimana OCEAN mendapat akurasi 67%, presisi 73%, recall 53%, dan F1Score 56%. Sedangkan dataset DISC mendapat akurasi 48%, presisi 37%, recall 30%, dan F1Score 24%.

Dari hasil penelitian yang dilakukan telah menunjukkan bahwa pemilihan dataset, pembagian rasio dataset, dan pemilihan feature extraction pada analisis profil menggunakan SVM dapat berdampak signifikan pada akurasi yang dihasilkan. Pada Gambar 4.13, ditunjukkan bahwa DISC menghasilkan akurasi yang lebih tinggi dibanding dataset OCEAN, namun jika melihat pada Gambar 4.14 ternyata dataset OCEAN masih dapat di optimalkan dengan mengubah rasio pembagian data dan nilai n-gram dari (1,2) menjadi (1,1). Sehingga, untuk

mendapatkan akurasi yang maksimal, sangat penting untuk memperhatikan pemilihan dataset, pembagian rasio dataset, dan feature extraction yang akan digunakan.

4.8. Analisis Perbandingan terhadap Studi Literature

Berdasarkan hasil analisis yang telah dilakukan pada dataset DISC dan OCEAN menggunakan SVM pada tweet yang berbahasa indonesia, selanjutnya akan ditampilkan rangkuman terkait studi literatur yang ada yang menjadi landasan pada penelitian ini. Rangkuman ini bertujuan untuk membandingkan hasil penelitian ini dengan hasil temuan studi literatur yang digunakan pada dataset yang sama-sama menggunakan Bahasa Indonesia dan dilakukan klasifikasi berdasarkan pengguna. Perbandingan hasil terhadap studi literatur dapat dilihat pada Tabel 4.13.

Tabel 4.13. Perbandingan dengan Penelitian Sebelumnya

Penulis	Dataset	Metode	Akurasi
(Utami et al., 2020)	DISC	KNN	28.33%
		Naïve Bayes	34.16%
Metode yang diusulakn	DISC	SVM (kernel linear)	56%
	OCEAN		67%

Berdasarkan perbandingan yang telah dilakukan dan perbedaan dalam hasil penelitian dalam Tabel 4.13, SVM dapat bekerja dengan baik untuk analisis profil pengguna twitter menggunakan tweet berbahasa Indonesia berdasarkan framework DISC maupun OCEAN. Hasil ini mengungguli hasil penelitian sebelumnya yang memiliki kemiripan pada bahasa yang digunakan pada dataset dan proses analisis profil berdasarkan pengguna (bukan berdasarkan *tweet*). Namun tetap perlu

diperhatikan bahwa walaupun dataset yang digunakan merupakan sama-sama *tweet* berbahasa Indonesia, data dalam perbandingan ini berasal dari dataset yang berbeda. Namun, walaupun menggunakan dataset yang berbeda, penelitian ini berhasil mengklasifikasikan kepribadian seseorang berdasarkan framework DISC dan OCEAN dengan akurasi yang cukup baik. Hasil ini menunjukkan kemajuan yang cukup dalam pengembangan model klasifikasi ini

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan dari hasil dan analisis penelitian yang telah dilakukan, dapat disimpulkan bahwa

1. SVM memiliki performa yang cukup baik dalam melakukan klasifikasi kepribadian menggunakan framework DISC dan OCEAN. Pada frameowrk DISC diperoleh akurasi 56%, presisi 63%, recall 36%, dan F1Score 34% dengan *preprocessing* data tanpa *stopword removal*, pembagian dataset dengan rasio 70:30, *feature extraction* n-gram, dan kernel SVM linear. Sedangkan pada dataset OCEAN diperoleh akurasi 67%, presisi 73%, recall 53%, dan F1Score 56% dengan *preprocessing* data tanpa *stopword removal*, pembagian dataset dengan rasio 60:40, *feature extraction* n-gram, dan kernel SVM linear.
2. Hasil pengujian pada penelitian ini menemukan bahwa penerapan *preprocessing*, pembagian data pemilihan *feature extraction*, dan pemilihan kernel SVM dapat mempengaruhi nilai akurasi, precision, recall, f1-score pada kedua framework dataset yang digunakan.
3. Adanya perbedaan yang tidak signifikan dari setiap skenario muncul karena kualitas dataset yang dari awal memang tidak banyak dan adanya ketidakseimbangan baik dilihat dari jumlah data setiap tweet maupun ketika sudah dikelompokkan untuk setiap user.

5.2. Saran

Berdasarkan analisis hasil percobaan yang dilakukan, terdapat beberapa saran yang dapat digunakan untuk penelitian selanjutnya, yaitu:

1. Menambahkan jumlah dan varian dataset lain baik dalam jumlah tweet untuk tiap user maupun jumlah user itu sendiri untuk mengoptimalkan keakuratan dan keandalan model dalam klasifikasi teks.
2. Melakukan eksplorasi lebih lanjut dalam melakukan optimasi dan meningkatkan akurasi seperti melakukan balancing data, menggunakan *feature extraction* yang lain, atau menggunakan model klasifikasi yang lain.

DAFTAR PUSTAKA

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

A Hartanto, E Utami, S Adi, H. H. (2019). 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 2019*, 533–536.

Ahmad, N., & Siddique, J. (2017). Personality Assessment using Twitter Tweets. *Procedia Computer Science*, *112*, 1964–1973. https://doi.org/10.1016/j.procs.2017.08.067

Artissa, Y. B. N. D., Asror, I., & Faraby, S. A. (2019). Personality Classification based on Facebook status text using Multinomial Naïve Bayes method. *Journal of Physics: Conference Series*, *1192*(1). https://doi.org/10.1088/1742-6596/1192/1/012003

Cernian, A., Vasile, N., & Sacala, I. S. (2021). Fostering cyber-physical social systems through an ontological approach to personality classification based on social media posts. *Sensors*, *21*(19). https://doi.org/10.3390/s21196611

Hootsuite. (2021). We-are-Social-Indonesian-Digital 2021. *Global Digital Insights*, 103.

Krasnova, H., Veltri, N. F., & Günther, O. (2012). Self-disclosure and Privacy Calculus on Social Networking Sites: The Role of Culture. *Business and Information Systems Engineering*, *4*(3), 127–135. https://doi.org/10.1007/s12599-012-0216-6

Langford, D., Fellows, R. F., Hancock, M. R., & Gale, A. W. (2020). Organizational behaviour. In *Human Resources Management in Construction*. https://doi.org/10.4324/9781315844695-9

Novák, J., Benda, P., Šilerová, E., Vaněk, J., & Kánská, E. (2021). Sentiment Analysis in Agriculture. *Agris On-Line Papers in Economics and Informatics*, *13*(1), 121–130. https://doi.org/10.7160/aol.2021.130109

Nurhayati, S. A., Nugrahawati, E. N., Dwarawati, D., Psikologi, P., & Psikologi, F. (2020). *Hubungan antara Personality dan Work Family Conflict Pada Karyawati Unisba*. https://doi.org/10.29313/.v6i2.23054

Ortigosa, A., Carro, R. M., & Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of Computer and System Sciences*, *80*(1), 57–71. https://doi.org/10.1016/j.jcss.2013.03.008

Ragab Bakry, M., Nasr, M. M., & Alsheref, F. K. (n.d.). Personality Classification Model of Social Network Profiles based on their Activities and Contents. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 13, Issue 7). www.ijacsa.thesai.org

Setiawan, H., & Wafi, A. A. (2020). Classification of Personality Type Based on Twitter Data Using Machine Learning Techniques. *2020 3rd International Conference on Information and Communications Technology, ICOIACT 2020*, 94–98. https://doi.org/10.1109/ICOIACT50329.2020.9332152

Somatdie, D. A., Widyarto, E., & Dwiyoga Widiantoro, A. (2019). Designing Web-Based DISC Psychology Personality Analysis Tests. *Journal of Information Systems*, *6*(2). https://doi.org/10.24167/Sisforma

Souri, A., Hosseinpour, S., & Rahmani, A. M. (2018). Personality classification based on profiles of social networks' users and the five-factor model of personality. In *Human-centric Computing and Information Sciences* (Vol. 8, Issue 1). https://doi.org/10.1186/s13673-018-0147-4

Tri Jaka, A. H. (2015). *Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining*.

Utami, E., Hartanto, A. D., Adi, S., Oyong, I., & Raharjo, S. (2022). Profiling analysis of DISC personality traits based on Twitter posts in Bahasa Indonesia. *Journal of King Saud University - Computer and Information Sciences*, *34*(2), 264–269. https://doi.org/10.1016/j.jksuci.2019.10.008

Utami, E., Raharjo, S., Dwi Hartanto, A., Adi, S., & Noor Ichsan, A. (2020, October 27). K-Nearest Neighbor and Naive Bayes Classifier Comparison for Individual Character Classification on Twitter. *2020 2nd International Conference on Cybernetics and Intelligent System, ICORIS 2020*. https://doi.org/10.1109/ICORIS50180.2020.9320759

Utami, N. A., Maharani, W., & Atastina, I. (2020). Personality Classification of Facebook Users According to Big Five Personality Using SVM (Support Vector Machine) Method. *Procedia Computer Science*, *196*, 348–355. https://doi.org/10.1016/j.procs.2021.12.023

Berliana, G., & Shaufiah, S. T. (2018). *Klasifikasi Posting Tweet mengenai Kebijakan Pemerintah Menggunakan Naive Bayesian Classification*.

Liang, H., Sun, X., Sun, Y., & Gao, Y. (2017). Text *feature extraction* based on deep learning: a review. In *Eurasip Journal on Wireless Communications and Networking* (Vol. 2017, Issue 1). Springer International Publishing. https://doi.org/10.1186/s13638-017-0993-1

Miftahuddin, Y., Pardede, J., & Andriani, A. A. (2016). *Perbandingan N-Gram Technique Dan Rabin Karp Pada Aplikasi Pendeteksi Plagiarisme Dokumen Teks Bahasa Indonesia*.