

# **Klasifikasi Kepribadian Big Five Berdasarkan Data Media Sosial X dengan Metode RoBERTa**

**Tugas Akhir**

**diajukan untuk memenuhi salah satu syarat memperoleh gelar sarjana  
dari Program Studi S1 Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**1301210333**

**Sachra Ramadhan**



**Universitas  
Telkom**

**Program Studi Sarjana Informatika**

**Fakultas Informatika**

**Universitas Telkom**

**Bandung**

**2025**

## LEMBAR PENGESAHAN

### **Klasifikasi Kepribadian Big Five Berdasarkan Data Media Sosial X dengan Metode RoBERTa**

#### *Big Five Personality Classification Based on Social Media X Data Using the RoBERTa Method*

**1301210333**

**Sachra Ramadhan**

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagai syarat  
memperoleh gelar pada Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung, 20 Januari 2025

Menyetujui

Pembimbing 1,



Dr. Warih Maharani, ST.,MT.

NIP : 01780020

Ketua Program Studi

Sarjana Informatika,



Dr. Erwin Budi Setiawan, S.Si., M.T.

NIP. 00760045

## LEMBAR ORISINALITAS

Dengan ini saya, Sachra Ramadhan, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul “Klasifikasi Kepribadian *Big Five* Berdasarkan Data Media Sosial X dengan Metode RoBERTa” berserta dengan seluruh isinya merupakan hasil karya saya sendiri, dengan tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang berlaku dengan masyarakat keilmuan, serta produk dari tugas akhir ini bukan merupakan hasil dari *Generative AI*. Saya siapanggung risiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam Laporan Tugas Akhir, atau jika ada klaim dari pihak lain terhadap keaslian karya.

Bandung, 20 Januari 2025

Yang menyatakan



Sachra Ramadhan

NIM 130121033

## ABSTRAK

Kepribadian mencerminkan pola pikir, emosi, dan perilaku individu, dan salah satu model yang digunakan untuk menganalisisnya adalah *Big Five Personality*, yang terdiri dari lima dimensi utama yaitu *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, dan *Neuroticism*. Penelitian ini bertujuan untuk mengklasifikasikan kepribadian *Big Five* menggunakan data dari media sosial X dengan metode RoBERTa. Kendala utama dalam penelitian ini adalah keterbatasan *dataset* yang hanya mencakup 381 pengguna serta ketidakseimbangan distribusi data antar label, yang memengaruhi akurasi prediksi.

Untuk mengatasi masalah ini, digunakan teknik *Random Oversampling* dan *Easy Data Augmentation* untuk menyeimbangkan dan memperkaya data. Model diuji dengan dua jenis *preprocessing*, yaitu *Half Preprocessing* dan *Full Preprocessing*, serta *tuning hyperparameter* menggunakan *Random Search*. Hasil eksperimen menunjukkan bahwa kombinasi *Data Augmentation* dan *Random Oversampling* memberikan performa terbaik, dengan akurasi 73,73% dan *F1-score* 73,37% pada skenario *Full Preprocessing*. Penelitian ini memberikan kontribusi dalam pengembangan metode klasifikasi berbasis teks untuk analisis kepribadian pada *dataset* kecil.

**Kata Kunci:** kepribadian *Big Five*, RoBERTa, *Random Oversampling*, *Data Augmentation*, media sosial, klasifikasi teks.

## ***ABSTRACT***

*Personality reflects an individual's patterns of thinking, emotions, and behavior, and one of the models used to analyze it is the Big Five Personality model, which consists of five main dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This study aims to classify Big Five personality traits using data from social media X with the RoBERTa method. The main challenge in this research is the limited dataset, which includes only 381 users, and the imbalance in data distribution across labels, which affects prediction accuracy.*

*To address this issue, Random Oversampling and Easy Data Augmentation techniques were applied to balance and enrich the data. The model was tested with two types of preprocessing: Half Preprocessing and Full Preprocessing, as well as hyperparameter tuning using Random Search. The experimental results show that the combination of Data Augmentation and Random Oversampling provides the best performance, with an accuracy of 73.73% and an F1-score of 73.37% in the Full Preprocessing scenario. This research contributes to the development of text-based classification methods for personality analysis on small datasets.*

**Keywords:** *Big Five Personality, RoBERTa, Random Oversampling, Data Augmentation, social media, text classification.*

## KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya, sehingga Tugas Akhir yang berjudul "Klasifikasi Kepribadian Big Five Berdasarkan Data Media Sosial X dengan Metode RoBERTa" dapat diselesaikan dengan baik.

Penelitian ini diharapkan mampu memberikan wawasan dan kontribusi yang bermanfaat dalam pengembangan prediksi berbasis teks, terutama dalam analisis data dari media sosial. Penulis menyadari bahwa Tugas Akhir ini masih memiliki berbagai keterbatasan dan kekurangan. Oleh sebab itu, kritik dan saran yang membangun sangat diharapkan sebagai bahan perbaikan di masa mendatang.

Akhir kata, semoga Tugas Akhir ini dapat memberikan manfaat bagi para pembaca serta menjadi kontribusi yang berarti dalam pengembangan ilmu pengetahuan dan teknologi.

Bandung, 20 Januari 2025

Penulis



Sachra Ramadhan

## UCAPAN TERIMA KASIH

Pada kesempatan ini, penulis ingin mengucapkan rasa terima kasih yang sebesar-besarnya kepada pihak-pihak yang telah memberikan dukungan, baik secara langsung maupun tidak langsung, selama penyusunan Tugas Akhir ini:

1. Ibu Dr. Warih Maharani, ST.,MT., selaku pembimbing, yang telah memberikan arahan, bimbingan, serta masukan yang sangat berarti selama proses penulisan Tugas Akhir ini.
2. Almarhum Bapak, yang telah memberikan doa, kasih sayang, dan dukungan yang tidak ternilai sepanjang hidup saya. Semoga Allah SWT memberikan tempat terbaik di sisi-Nya.
3. Ibu dan keluarga, atas doa, kasih sayang, serta dukungan moral dan material yang selalu menguatkan penulis dalam menjalani setiap langkah perjalanan akademik ini.
4. Teman-teman seperjuangan, yang telah memberikan semangat, bantuan, dan kebersamaan yang sangat berarti selama proses penelitian dan penulisan Tugas Akhir ini.
5. Semua pihak lainnya yang tidak dapat disebutkan satu per satu, namun telah memberikan kontribusi dalam berbagai bentuk selama penyusunan Tugas Akhir ini.

Semoga segala bantuan, doa, dan dukungan yang diberikan kepada penulis mendapatkan balasan yang terbaik dari Tuhan Yang Maha Esa. Akhir kata, semoga Tugas Akhir ini dapat bermanfaat bagi pembaca dan menjadi inspirasi untuk penelitian lebih lanjut.

# DAFTAR ISI

LEMBAR PENGESAHAN .....	ii
LEMBAR ORISINALITAS .....	iii
ABSTRAK .....	iv
<i>ABSTRACT</i> .....	v
KATA PENGANTAR.....	vi
UCAPAN TERIMA KASIH .....	vii
DAFTAR ISI.....	viii
DAFTAR GAMBAR.....	xi
DAFTAR TABEL .....	xii
<b>BAB 1 PENDAHULUAN .....</b>	<b>13</b>
1.1. Latar Belakang .....	13
1.2. Rumusan Masalah.....	15
1.3. Tujuan dan Manfaat .....	15
1.4. Batasan Masalah .....	16
1.5. Metode Penelitian .....	16
1.5.1. Studi Literatur .....	17
1.5.2. Pengumpulan Data .....	17
1.5.3. <i>Preprocessing Data</i> .....	17
1.5.4. <i>Tuning Hyperparameter</i> .....	18
1.5.5. Pengembangan Model .....	18
1.5.6. Eksperimen dan Evaluasi .....	20
<b>BAB 2 TINJAUAN PUSTAKA.....</b>	<b>21</b>
2.1. Penelitian Terkait .....	21
2.2. Media Sosial.....	23
2.3. X.....	23
2.4. Klasifikasi Kepribadian.....	23
2.5. RoBERTa .....	24
2.5.1. <i>Embedding Layer</i> .....	24
2.5.2. <i>Pre-trained Model</i> .....	25
2.5.3. <i>Dynamic Masking</i> .....	25
2.5.4. <i>Fine Tuning</i> .....	25
2.6. <i>Big Five Personality</i> .....	25
2.7. <i>Random Oversampling (ROS)</i> .....	26

2.8.	<i>Data Augmentation</i> .....	26
2.9.	Matriks Evaluasi .....	27
<b>BAB 3</b>	<b>PERANCANGAN SISTEM</b> .....	<b>30</b>
3.1.	Desain Perancangan Sistem.....	30
3.2.	Pengambilan Data .....	31
3.3.	Pelabelan Data.....	31
3.4.	<i>Preprocessing Data</i> .....	32
3.4.1.	<i>Cleansing</i> .....	32
3.4.2.	<i>Case Folding</i> .....	33
3.4.3.	<i>Tokenization</i> .....	33
3.4.4.	<i>Normalization</i> .....	33
3.4.5.	<i>Stopword</i> .....	34
3.4.6.	<i>Stemming</i> .....	34
3.4.7.	<i>Half Preprocessing</i> .....	34
3.4.8.	<i>Full Preprocessing</i> .....	35
3.5.	<i>Hyperparameter Tuning</i> .....	36
3.6.	Pemodelan Roberta .....	37
3.7.	Pendekatan Model.....	38
3.7.1.	<i>Baseline</i> .....	38
3.7.2.	<i>Data Augmentation</i> .....	38
3.7.3.	<i>Random Oversampling (ROS)</i> .....	39
3.7.4.	<i>ROS + Data Augmentation</i> .....	39
3.8.	Evaluasi .....	39
<b>BAB 4</b>	<b>HASIL PERCOBAAN DAN ANALISIS</b> .....	<b>41</b>
4.1.	<i>Dataset</i> .....	41
4.2.	<i>Hyperparameter Tuning</i> .....	42
4.3.	Skenario Percobaan .....	44
4.3.1.	Skenario 1 : <i>Baseline</i> .....	45
4.3.2.	Skenario 2 : <i>Data Augmentation</i> .....	46
4.3.3.	Skenario 3 : <i>Random Oversampling (ROS)</i> .....	47
4.3.4.	Skenario 4 : <i>Data Augmentation + ROS</i> .....	49
4.4.	Hasil Percobaan.....	50
4.4.1.	Hasil Percobaan 1 : <i>Baseline</i> .....	50
4.4.2.	Hasil Percobaan 2 : <i>Data Augmentation</i> .....	54
4.4.3.	Hasil Percobaan 3 : <i>Random Oversampling (ROS)</i> .....	58
4.4.4.	Hasil Percobaan 4 : <i>Data Augmentation + ROS</i> .....	62
4.5.	Analisis .....	66
4.5.1.	Analisis untuk <i>Full Preprocessing</i> .....	67

4.5.2.	Analisis untuk <i>Half Preprocessing</i> .....	68
4.5.3.	Perbandingan Kinerja antara <i>Data Split</i> .....	68
<b>BAB 5</b>	<b>KESIMPULAN DAN SARAN</b> .....	<b>70</b>
5.1.	Kesimpulan .....	70
5.2.	Saran .....	70
<b>DAFTAR PUSTAKA</b>	.....	<b>71</b>

## DAFTAR GAMBAR

Gambar 2. 1 Embedding Layer pada RoBERTa [19].....	24
Gambar 3. 1 Alur Diagram Sistem.....	30
Gambar 3. 2 <i>Code</i> Inisiasi <i>Trainer</i> Model RoBERTa .....	37
Gambar 4. 1 Distribusi Label <i>Dataset</i> .....	41
Gambar 4. 2 Distribusi Label <i>Baseline</i> .....	45
Gambar 4. 3 Distribusi Label Setelah <i>Data Augmentation</i> .....	47
Gambar 4. 4 Distribusi Label Setelah <i>Random Oversampling</i> (ROS).....	48
Gambar 4. 5 Distribusi Label Setelah <i>Data Augmentation</i> + ROS.....	49
Gambar 4. 6 <i>Confusion Matrix</i> untuk <i>Baseline</i> dengan <i>Full Preprocessing</i> (80/20) .....	51
Gambar 4. 7 <i>Confusion Matrix</i> untuk <i>Baseline</i> dengan <i>Full Preprocessing</i> (70/30) .....	52
Gambar 4. 8 <i>Confusion Matrix</i> untuk <i>Baseline</i> dengan <i>Half Preprocessing</i> (80/20) .....	52
Gambar 4. 9 <i>Confusion Matrix</i> untuk <i>Baseline</i> dengan <i>Half Preprocessing</i> (70/30) .....	53
Gambar 4. 10 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> dengan <i>Full Preprocessing</i> (80/20) .....	55
Gambar 4. 11 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> dengan <i>Full Preprocessing</i> (70/30) .....	56
Gambar 4. 12 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> dengan <i>Half Preprocessing</i> (80/20) .....	56
Gambar 4. 13 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> dengan <i>Half Preprocessing</i> (70/30) .....	57
Gambar 4. 14 <i>Confusion Matrix</i> untuk <i>Random Oversampling</i> dengan <i>Full Preprocessing</i> (80/20) .....	59
Gambar 4. 15 <i>Confusion Matrix</i> untuk <i>Random Oversampling</i> dengan <i>Full Preprocessing</i> (70/30) .....	60
Gambar 4. 16 <i>Confusion Matrix</i> untuk <i>Random Oversampling</i> dengan <i>Half Preprocessing</i> (80/20) .....	60
Gambar 4. 17 <i>Confusion Matrix</i> untuk <i>Random Oversampling</i> dengan <i>Half Preprocessing</i> (70/30) .....	61
Gambar 4. 18 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> + ROS dengan <i>Full Preprocessing</i> (80/20) .....	63
Gambar 4. 19 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> + ROS dengan <i>Full Preprocessing</i> (70/30) .....	64
Gambar 4. 20 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> + ROS dengan <i>Half Preprocessing</i> (80/20) .....	64
Gambar 4. 21 <i>Confusion Matrix</i> untuk <i>Data Augmentation</i> + ROS dengan <i>Half Preprocessing</i> (70/30) .....	65

## DAFTAR TABEL

Tabel 2. 1 Penelitian Terkait.....	22
Tabel 2. 2 Contoh Penerapan <i>Easy Data Augmentation</i> .....	27
Tabel 3. 1 Pelabelan Data.....	31
Tabel 3. 2 <i>Cleansing</i> .....	32
Tabel 3. 3 <i>Case Folding</i> .....	33
Tabel 3. 4 <i>Tokenization</i> .....	33
Tabel 3. 5 <i>Normalization</i> .....	33
Tabel 3. 6 <i>Stopword</i> .....	34
Tabel 3. 7 <i>Stemming</i> .....	34
Tabel 3. 8 Hasil <i>Half Preprocessing</i> .....	35
Tabel 3. 9 Hasil <i>Full Preprocessing</i> .....	36
Tabel 3. 10 Contoh nilai evaluasi .....	40
Tabel 4. 1 Hasil Percobaan Skenario 1 <i>Baseline</i> .....	51
Tabel 4. 2 Hasil Percobaan Skenario 2 <i>Data Augmentation</i> .....	55
Tabel 4. 3 Hasil Percobaan Skenario 3 <i>Random Oversampling (ROS)</i> .....	59
Tabel 4. 4 Hasil Percobaan Skenario 4 <i>Data Augmentation + ROS</i> .....	63
Tabel 4. 5 Hasil Percobaan Uji Skenario .....	66

# BAB 1 PENDAHULUAN

## 1.1. Latar Belakang

Media sosial telah menjadi bagian penting dalam kehidupan sehari-hari, termasuk di Indonesia yang merupakan salah satu negara dengan jumlah pengguna X (Twitter) terbesar di dunia, dengan sekitar 27,5 juta pengguna aktif pada Oktober 2023 [1]. X, sebagai platform media sosial populer, digunakan oleh banyak orang di berbagai negara untuk berbagi pemikiran dan aktivitas melalui kalimat-kalimat singkat. Basis pengguna yang besar dan aktif ini menjadikan X tidak hanya sebagai sumber informasi yang relevan, tetapi juga sebagai peluang untuk menganalisis perilaku individu, termasuk pola kepribadian, melalui penelitian berbasis teks [16]. Meskipun data dari media sosial sangat melimpah, tantangan utama yang dihadapi adalah bagaimana mengekstrak informasi kepribadian dengan akurat dari teks yang bersifat tidak terstruktur dan tidak merata.

Kepribadian mencakup pola pikir, emosi, dan perilaku yang membedakan setiap individu [2]. Salah satu model yang sering digunakan untuk menganalisisnya adalah *Big Five Personality* (OCEAN), yang terdiri dari lima dimensi: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, dan *Neuroticism* [3]. Model ini menawarkan cara yang komprehensif untuk memahami kepribadian dan banyak diterapkan dalam psikologi. Untuk mengukur dimensi-dimensi tersebut, biasanya digunakan kuesioner seperti BFI-44, yang diperkenalkan oleh John, Donahue, dan Kentle pada tahun 1991. BFI-44 mencakup 44 pertanyaan, di mana peserta menanggapi setiap pernyataan menggunakan skala Likert dari "sangat tidak setuju" hingga "sangat setuju", untuk menunjukkan seberapa relevan pernyataan tersebut dengan diri mereka. Kuesioner ini telah terbukti valid di berbagai kelompok budaya, menjadikannya alat yang populer dan efektif untuk menilai kepribadian [20], [21].

Namun, penerapan model *Big Five Personality* dalam analisis data media sosial menghadapi beberapa kendala. Salah satu tantangan utama dalam penelitian ini adalah keterbatasan jumlah data pengguna, yaitu hanya 381 akun dengan cuitan yang telah diberi label kepribadian. Selain jumlahnya yang kecil, distribusi label

dalam *dataset* juga tidak seimbang, di mana kelas mayoritas lebih dominan dalam prediksi, sementara kelas minoritas kurang terwakili. Ketidakseimbangan ini dapat menyebabkan bias dalam hasil klasifikasi, mengurangi akurasi model, dan membatasi kemampuan model dalam melakukan generalisasi dengan baik [22].

Untuk mengatasi tantangan ini, penelitian ini mengusulkan penggunaan dua teknik utama yaitu *Random Oversampling* (ROS) dan *Easy Data Augmentation* (EDA). ROS bertujuan untuk menyeimbangkan distribusi kelas dengan menggandakan sampel dari kelas minoritas, sehingga model dapat belajar secara lebih adil terhadap semua kelas. Sementara itu, EDA memperkaya dataset dengan teknik augmentasi teks seperti penggantian sinonim, penyisipan acak, pertukaran kata, dan penghapusan kata. Dengan menerapkan kedua teknik ini, diharapkan model dapat memiliki data yang lebih representatif dan meningkatkan akurasi klasifikasi [14], [15].

Sebagai model utama dalam penelitian ini, digunakan *A Robustly Optimized BERT Pretraining Approach* (RoBERTa) untuk mengklasifikasikan kepribadian *Big Five* berdasarkan data dari X. RoBERTa merupakan pengembangan dari *Bidirectional Encoder Representations from Transformers* (BERT) yang menghilangkan *Next Sentence Prediction* (NSP) dan menerapkan *dynamic masking* untuk meningkatkan variasi data yang dilihat oleh model. Dengan prosedur *pretraining* yang lebih optimal dan data yang lebih besar, RoBERTa telah terbukti unggul dalam berbagai tugas pemrosesan bahasa alami, termasuk analisis sentimen dan klasifikasi kepribadian [4].

Penelitian sebelumnya menunjukkan bahwa RoBERTa mampu melakukan tugas analisis sentimen dan identifikasi kepribadian. Beberapa contoh penerapannya termasuk penelitian oleh Muhammad Mahrus Zain et al. [5] dan Eggi Farkhan Tsani serta Derwin Suhartono [6], yang memanfaatkan RoBERTa dalam analisis sentimen dan klasifikasi kepribadian. Selain itu, penelitian lain seperti yang dilakukan oleh Yani dan Maharani [7] yang menggunakan RoBERTa untuk menganalisis konten *cyberbullying* di X, serta Putra dan Setiawan [8] yang mengaplikasikan RoBERTa dalam analisis sentimen, juga menunjukkan efektivitas metode ini dalam memproses data media sosial. Penelitian oleh Murarka et al. [9]

mengenai deteksi penyakit mental lebih lanjut membuktikan keberhasilan RoBERTa dalam menghadapi tugas klasifikasi teks yang kompleks.

Dengan menerapkan kombinasi RoBERTa, ROS, dan EDA, penelitian ini bertujuan untuk meningkatkan akurasi dalam klasifikasi kepribadian *Big Five* berbasis data media sosial X. Pendekatan ini diharapkan dapat mengatasi tantangan ketidakseimbangan data serta memberikan kontribusi dalam pengembangan metode klasifikasi teks berbasis kepribadian menggunakan kecerdasan buatan.

## 1.2. Rumusan Masalah

Dalam penelitian ini, terdapat beberapa permasalahan yang dirumuskan sebagai berikut:

1. Bagaimana metode RoBERTa dapat diimplementasikan untuk mengklasifikasikan kepribadian *Big Five* berdasarkan data dari media sosial X?
2. Seberapa baik performansi metode RoBERTa dalam klasifikasi kepribadian *Big Five* berdasarkan metrik evaluasi yang meliputi akurasi, *F1-score*, presisi, dan *recall*?
3. Pendekatan apa yang paling efektif untuk meningkatkan performansi model dalam klasifikasi kepribadian *Big Five* dengan menggunakan metode RoBERTa?

## 1.3. Tujuan dan Manfaat

Tujuan dari penelitian ini adalah sebagai berikut:

1. Mengimplementasikan metode RoBERTa dalam klasifikasi kepribadian *Big Five* berdasarkan data dari media sosial X.
2. Mengevaluasi performansi RoBERTa melalui metrik evaluasi seperti akurasi, *F1-score*, presisi, dan *recall*.
3. Mengidentifikasi pendekatan yang paling efektif dalam meningkatkan performansi model dalam klasifikasi kepribadian *Big Five*.

Sementara itu, manfaat dari penelitian ini diharapkan dapat memberikan kontribusi

sebagai referensi untuk penelitian lebih lanjut di bidang klasifikasi teks, khususnya dalam konteks klasifikasi multi-label untuk klasifikasi kepribadian.

#### 1.4. Batasan Masalah

Penelitian ini memiliki beberapa batasan untuk menyederhanakan ruang lingkup dan memastikan penyelesaian penelitian dalam waktu yang tersedia. Batasan-batasan tersebut adalah sebagai berikut:

1. Data diperoleh melalui *crawling* cuitan dari pengguna media sosial X yang telah mengisi kuesioner *Big Five Inventory* (BFI-44). Hasil kuesioner ini digunakan sebagai label untuk lima dimensi *Big Five*, dengan total data yang terdiri dari 381 akun. Pemilihan jumlah akun ini didasarkan pada ketersediaan data yang telah mengisi kuesioner BFI-44 yang lengkap, tanpa adanya pembatasan jumlah cuitan yang diambil.
2. Sebagian besar cuitan dalam *dataset* ditulis dalam bahasa Indonesia, mengingat fokus penelitian pada pengguna media sosial X di Indonesia. Namun, beberapa cuitan mungkin menggunakan bahasa Inggris.
3. Penelitian ini mencakup lima dimensi kepribadian dari model *Big Five Personality: openness, conscientiousness, extraversion, agreeableness, dan neuroticism*.
4. Penelitian ini menggunakan metode RoBERTa sebagai model utama dalam klasifikasi.
5. Penelitian ini terbatas pada penggunaan *Random Oversampling* (ROS) dan *Easy Data Augmentation* (EDA) sebagai teknik untuk menangani ketidakseimbangan data dan meningkatkan performa model.
6. Penelitian ini hanya mengukur kepribadian yang nampak dari cuitan pengguna saat ini.

#### 1.5. Metode Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang terstruktur, meliputi studi literatur, pengumpulan data, *preprocessing data*, pengembangan model, serta

eksperimen dan evaluasi. Setiap tahapan dirancang untuk meningkatkan akurasi model dan memastikan pendekatan yang efektif.

### **1.5.1. Studi Literatur**

Penelitian ini diawali dengan mempelajari teori *Big Five Personality*, metode RoBERTa, dan teknik pengolahan data teks untuk klasifikasi multi-label. Literatur yang digunakan terdiri dari jurnal, buku, dan artikel ilmiah yang relevan, yang mendukung dasar teori dan pendekatan yang digunakan dalam penelitian ini.

### **1.5.2. Pengumpulan Data**

Data diperoleh melalui *crawling* cuitan dari pengguna media sosial X yang telah mengisi kuesioner *Big Five Inventory* (BFI-44). Hasil kuesioner ini digunakan sebagai label untuk lima dimensi kepribadian *Big Five*: *openness*, *conscientiousness*, *extraversion*, *agreeableness*, dan *neuroticism*. *Dataset* yang dikumpulkan mencakup cuitan dalam bahasa Indonesia dan sebagian dalam bahasa Inggris, yang menggambarkan kebiasaan pengguna media sosial di Indonesia.

Penelitian ini mengambil seluruh cuitan dari 381 akun pengguna yang telah mengisi kuesioner BFI-44, tanpa membatasi jumlah cuitan yang diambil. Namun, distribusi cuitan antar akun tidak merata, yang dapat mempengaruhi akurasi model. Akun dengan banyak cuitan cenderung lebih dominan, sementara akun dengan sedikit cuitan mungkin kurang terwakili. Untuk mengatasi masalah ini, digunakan Random Oversampling (ROS) dan Data Augmentation (EDA) untuk menyeimbangkan data dan memperkaya variasinya. Meskipun demikian, ketidakseimbangan ini tetap menjadi batasan yang perlu dipertimbangkan saat menilai hasil penelitian.

### **1.5.3. Preprocessing Data**

Proses *preprocessing* bertujuan untuk memastikan data memiliki kualitas tinggi, meminimalkan *noise*, dan meningkatkan efisiensi pelatihan model. Penelitian ini menggunakan dua skenario *preprocessing*:

1. *Half Preprocessing*: Meliputi *cleansing*, *case folding*, dan *tokenization*. Pada skenario ini, data hanya dibersihkan dan distandarisasi, serta diproses menjadi token.

2. *Full Preprocessing*: Meliputi *cleansing*, *case folding*, *tokenization*, *normalization*, penghapusan *stopword*, dan *stemming*. Skenario ini lebih komprehensif dan bertujuan untuk mengurangi *noise* dalam data yang dapat mengganggu proses pelatihan.

Kedua skenario preprocessing ini diuji untuk melihat pengaruh kompleksitas dan teknik preprocessing terhadap akurasi model RoBERTa. Pengujian ini bertujuan untuk mengidentifikasi proses preprocessing yang memberikan kontribusi terbesar dalam meningkatkan performansi model dalam klasifikasi kepribadian berdasarkan *Big Five*.

#### 1.5.4. *Tuning Hyperparameter*

Setelah tahap preprocessing, dilakukan pencarian *hyperparameter* terbaik menggunakan metode *random search*. Parameter yang disesuaikan meliputi *learning rate*, *batch size*, dan *weight decay*, yang mempengaruhi performa model saat pelatihan.

Metode *random search* dipilih karena memungkinkan pencarian kombinasi parameter yang lebih efisien dan fleksibel. Proses *tuning* dilakukan terpisah pada *dataset* dengan *half preprocessing* dan *full preprocessing*, untuk menyesuaikan parameter dengan karakteristik masing-masing *dataset*. Kombinasi parameter terbaik kemudian digunakan untuk melatih model utama.

#### 1.5.5. **Pengembangan Model**

Pengembangan model dilakukan menggunakan RoBERTa sebagai model utama untuk klasifikasi kepribadian *Big Five*. Beberapa pendekatan diterapkan selama pelatihan untuk meningkatkan performa model, antara lain:

1. *Baseline*: Model dilatih tanpa penyeimbangan data, sebagai acuan performa awal.
2. *Random Oversampling (ROS)*: Teknik ini menyeimbangkan distribusi data antar label dengan menduplikasi data dari label minoritas hingga setara dengan mayoritas [14].

3. *Data Augmentation*: Teknik augmentasi data dilakukan untuk menambah variasi *dataset*, menggunakan metode *Easy Data Augmentation* (EDA) yang mencakup:

- *Synonym Replacement* (SR): Mengganti kata-kata dalam teks dengan sinonimnya secara acak, untuk menambah keragaman kalimat.
- *Random Insertion* (RI): Menyisipkan kata acak ke dalam teks untuk menambah variasi.
- *Random Swap* (RS): Menukar posisi dua kata secara acak dalam kalimat untuk menciptakan variasi struktur tanpa merubah makna.
- *Random Deletion* (RD): Menghapus kata-kata dalam teks dengan probabilitas tertentu, bertujuan untuk memperkenalkan variasi dalam panjang teks dan mengurangi ketergantungan model pada kata-kata tertentu.

Penerapan teknik ini bertujuan untuk memperkaya data dan membantu model mengenali pola yang lebih beragam.

4. *Data Augmentation* + ROS: Kombinasi antara *Data Augmentation* dan *Random Oversampling* dilakukan untuk menciptakan *dataset* yang lebih beragam sekaligus seimbang.

5. Pengujian Pembagian Data

Setiap pendekatan diterapkan pada dua skenario pembagian data:

- 80/20: 80% data digunakan untuk pelatihan dan 20% untuk pengujian.
- 70/30: 70% data digunakan untuk pelatihan dan 30% untuk pengujian.

Dengan kombinasi teknik yang berbeda, enam belas skenario pengujian diterapkan untuk mengidentifikasi pendekatan terbaik dalam meningkatkan performa model klasifikasi multi-label, khususnya dalam prediksi kepribadian *Big Five*.

### **1.5.6. Eksperimen dan Evaluasi**

Eksperimen dilakukan untuk mengevaluasi kombinasi *preprocessing* dan pendekatan yang diterapkan. Evaluasi model dilakukan menggunakan metrik akurasi, *F1-score*, presisi, dan *recall*, dengan tujuan untuk mengidentifikasi pendekatan terbaik dalam meningkatkan performa model.

## **BAB 2 TINJAUAN PUSTAKA**

### **2.1. Penelitian Terkait**

Dalam penelitian ini, terdapat sejumlah referensi dari beberapa penelitian sebelumnya yang berkaitan dengan latar belakang tugas akhir ini. Penelitian-penelitian yang relevan terlampir pada Tabel 2.1.

Tabel 2. 1 Penelitian Terkait

Judul	Metode	Tingkat Kinerja
Analisis Sentimen Pendapat Masyarakat Mengenai Vaksin Covid-19 Pada Media Sosial Twitter dengan <i>Robustly Optimized BERT Pretraining Approach</i> [5]	RoBERTa	<ul style="list-style-type: none"> <li>Rata-rata akurasi keseluruhan data sebesar 95%.</li> <li>Akurasi rata-rata prediksi untuk label positif 84%, label netral 97%, dan label negatif 93%.</li> </ul>
<i>Personality Identification from Social Media Using Ensemble BERT and RoBERTa</i> [6]	Ensemble Bert, RoBERTa dan <i>Big Five</i> sebagai model kepribadian	<ul style="list-style-type: none"> <li><i>Dataset</i> Twitter mendapatkan rata-rata <i>F1 score</i> sebesar 0,730.</li> <li><i>Dataset</i> Youtube mendapatkan rata-rata <i>F1 score</i> sebesar 0,741.</li> </ul>
<i>Analyzing Cyberbullying Negative Content on Twitter Social Media With The RoBERTa Method</i> [7]	RoBERTa	<ul style="list-style-type: none"> <li><i>Pre-processing</i> mendapat 86,9% akurasi dan 77,5% <i>F1 score</i>.</li> <li><i>Fullprocessing</i> mendapat 86,9% akurasi dan 75,7% <i>F1 score</i>.</li> </ul>
<i>Sentiment Analysis on Social Media with Glove Using CNN and RoBERTa</i> [8]	CNN, TF-IDF, RoBERTa, SMOTE, Glove	<ul style="list-style-type: none"> <li>Akurasi TF-IDF dan RoBERTa sebesar 87,75%.</li> <li>Baseline unigram menggunakan CNN mendapat akurasi 86,79%.</li> <li>Akurasi baseline unigram, RoBERTa, SMOTE, dan cuitan korpus Top10 menghasilkan akurasi tertinggi sebesar 95,56%.</li> </ul>
<i>Detection and Classification of mental illnesses on social media using RoBERTa</i> [9]	RoBERTa	<ul style="list-style-type: none"> <li>Mendapat <i>F1 score</i> sebesar 0,82 dengan 10% pengujian yang dimodifikasi.</li> <li>Mendapat <i>F1 score</i> 0,71 dengan 50% pengujian yang dimodifikasi.</li> </ul>

## 2.2. Media Sosial

Media sosial telah menjadi kebiasaan dalam kehidupan sehari-hari dan telah berkembang dengan cepat di seluruh dunia. Banyak orang menggunakannya untuk berkomunikasi dan berbagi ide, diskusi, dan pendapat [10], [17]. Tercatat pada tahun 2025 ada sebanyak 5,22 miliar pengguna sosial media secara global dan sebanyak 228,76 juta pengguna sosial media dari Indonesia pada tahun 2023 dan akan terus meningkat tiap tahunnya<sup>1</sup>. Media sosial juga menjadi sumber data yang penting dalam penelitian-penelitian seperti analisis sentimen dan analisis kepribadian.

## 2.3. X

X adalah platform media sosial yang memungkinkan penggunanya untuk mengirim pesan teks yang disebut cuitan, dengan panjang pesan hingga 280 karakter. X memiliki pengguna yang sangat besar dan aktif, menjadikannya sebagai sumber data yang relevan untuk penelitian, termasuk dalam analisis kepribadian berbasis teks dari cuitan yang diposting oleh penggunanya [17], [18].

## 2.4. Klasifikasi Kepribadian

Klasifikasi kepribadian sangat penting dalam berbagai industri. Di bidang pemasaran, data kepribadian dapat digunakan untuk menyesuaikan strategi pemasaran. Dalam manajemen sumber daya manusia, pemahaman tentang kepribadian seseorang dapat membantu dalam penempatan pekerjaan yang tepat dan meningkatkan kinerja tim [2]. Dalam penelitian ini, digunakan model *Big Five Personality* untuk mengklasifikasikan kepribadian pengguna media sosial.

---

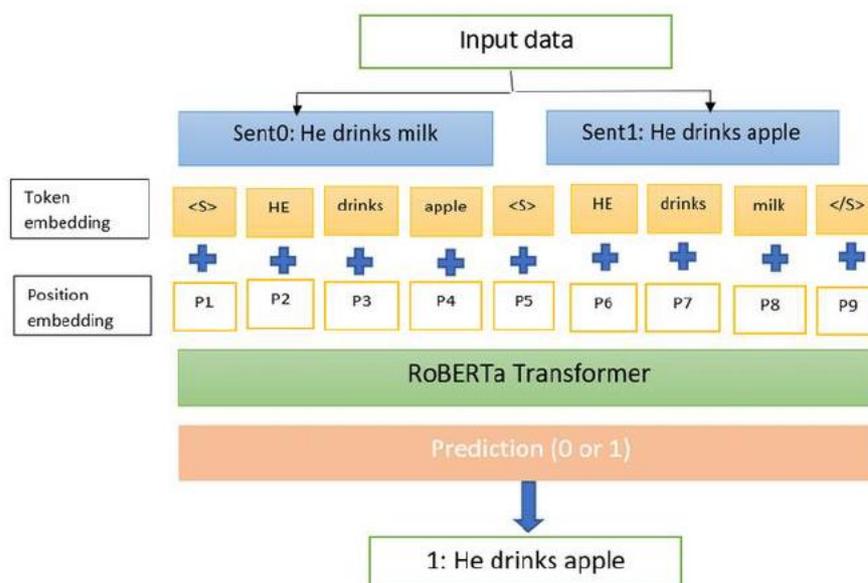
<sup>1</sup> Shubham Singh, "Social Media User 2025 Statistic" (<https://www.demandsage.com/social-media-users/>), Diakses pada 27 Desember 2024,2024

## 2.5. RoBERTa

RoBERTa (*Robustly Optimized BERT Pretraining Approach*) adalah model berbasis transformer yang dikembangkan untuk meningkatkan kinerja BERT (*Bidirectional Encoder Representations from Transformers*). RoBERTa memodifikasi beberapa aspek BERT, termasuk penghapusan tugas *Next Sentence Prediction* (NSP) dan penggunaan teknik *dynamic masking*. Dengan menggunakan *dataset* yang lebih besar dan durasi pelatihan yang lebih lama, RoBERTa menunjukkan kinerja yang lebih baik untuk tugas klasifikasi teks seperti analisis sentimen dan klasifikasi kepribadian [4].

### 2.5.1. Embedding Layer

RoBERTa menggabungkan *token embedding* dan *position embedding* untuk mengonversi kata menjadi vektor numerik, sekaligus memberikan informasi mengenai posisi kata dalam kalimat. Hal ini memungkinkan model untuk lebih baik dalam memahami konteks antar kata. Terlihat pada Gambar 2.1, kedua jenis *embedding* tersebut digabungkan sebelum diproses lebih lanjut oleh RoBERTa Transformer. Dengan menerapkan teknik *Masked Language Modeling* (MLM) yang lebih efektif, RoBERTa menghasilkan representasi teks yang lebih kuat, mendukung berbagai aplikasi NLP dengan kinerja yang lebih optimal [19].



Gambar 2. 1 Embedding Layer pada RoBERTa [19]

### **2.5.2. Pre-trained Model**

RoBERTa dilatih pada *dataset* yang jauh lebih besar dibandingkan BERT, seperti BookCorpus, Wikipedia, dan CC-News. Ini memberikan RoBERTa akses ke beragam teks, memungkinkan model untuk memahami berbagai konteks dan meningkatkan kemampuan generalisasinya untuk tugas-tugas pemrosesan bahasa alami [4].

### **2.5.3. Dynamic Masking**

RoBERTa menggantikan teknik NSP yang digunakan pada BERT dengan *dynamic masking*. Teknik ini mengganti token secara acak setiap kali model memproses data, bukan hanya sekali selama pelatihan. Pendekatan ini memperbaiki efisiensi dan memungkinkan model untuk mempelajari lebih banyak variasi dalam pola bahasa, yang berkontribusi pada peningkatan kinerja model, terutama pada tugas yang memerlukan pemahaman konteks yang lebih mendalam [4].

### **2.5.4. Fine Tuning**

Setelah *pre-training*, RoBERTa dapat di *fine-tune* untuk tugas spesifik dengan menggunakan *dataset* yang lebih terfokus dan sesuai dengan konteks tugas tersebut. Proses *fine-tuning* melibatkan penyesuaian *hyperparameter* seperti *learning rate*, *batch size*, dan *weight decay* untuk mengoptimalkan performa model. Pengaturan *hyperparameter* yang tepat memainkan peran kunci dalam *fine-tuning*, yang memungkinkan RoBERTa untuk beradaptasi dengan baik terhadap tugas-tugas seperti analisis sentimen dan pemahaman teks [4].

## **2.6. Big Five Personality**

Model kepribadian *Big Five Personality* terdiri dari lima dimensi, dengan masing-masing menggambarkan ciri kepribadian utama seseorang. Faktor-faktor ini termasuk *Openness*, *Conscientiousness*, *Extrovert*, *Agreeableness*, dan *Neuroticism*. *Openness* menunjukkan orang yang sangat terbuka, ingin tahu, dan perhatian yang tinggi, sedangkan *Conscientiousness* menunjukkan orang yang cermat, berhati-hati, dan teliti. *Extrovert* memiliki sifat ramah, mudah bergaul, ceria, dan antusias, sedangkan *Neuroticism* cemas, gugup, ragu-ragu, dan mudah

frustas, lalu terakhir *Agreeableness* cenderung empati, perhatian, dan memiliki dorong yang kuat untuk bekerja sama dalam tim [11].

Model ini diukur menggunakan skala BFI-44, yang dikembangkan oleh John, Donahue, dan Kentle pada tahun 1991 [21]. Skala ini terdiri dari 44 pertanyaan yang melakukan analisis lima dimensi yang ada pada *Big Five Personality*. Skala ini dapat menguji kepribadian secara subjek secara akurat, tidak seperti skala dengan jumlah pertanyaan yang terlalu sedikit, skala ini dapat juga dapat secara akurat menguji karakteristik kepribadian dari subjek [12].

## **2.7. *Random Oversampling (ROS)***

*Random Oversampling (ROS)* adalah teknik yang digunakan untuk mengatasi ketidakseimbangan kelas dalam *dataset*. Dalam ROS, sampel dari kelas minoritas digandakan hingga jumlahnya seimbang dengan kelas mayoritas. Teknik ini membantu model belajar lebih baik dari data minoritas dan menghindari bias yang mungkin terjadi jika kelas mayoritas lebih dominan [14].

## **2.8. *Data Augmentation***

*Data augmentation* adalah teknik untuk menambah variasi data dengan melakukan perubahan kecil pada data asli. Teknik ini sangat berguna untuk mengatasi masalah *dataset* kecil dan tidak seimbang, karena bisa memperkaya data tanpa perlu menambah data baru dari sumber lain. Dalam penelitian ini, *Easy Data Augmentation (EDA)* digunakan dengan empat teknik utama yang bertujuan untuk meningkatkan variasi kalimat, membantu model lebih generalisasi, dan mengurangi *overfitting*. Berikut contoh dari masing-masing teknik:

Tabel 2. 2 Contoh Penerapan *Easy Data Augmentation*

<b>Teknik</b>	<b>Contoh Kalimat Asli</b>	<b>Contoh Setelah Data Augmentation</b>	<b>Deskripsi</b>
<i>Synonym Replacement</i>	"Saya sangat senang hari ini."	"Saya sangat bahagia hari ini."	Mengganti kata dengan sinonim, memperkaya variasi kalimat tanpa mengubah makna.
<i>Random Insertion</i>	"Saya akan pergi ke pasar."	"Saya akan besok pergi ke pasar."	Menambah kata acak untuk memperkenalkan variasi dalam kalimat.
<i>Random Swap</i>	"Saya membeli buku baru."	"Saya buku membeli baru."	Menukar posisi kata untuk menciptakan variasi struktur kalimat.
<i>Random Deletion</i>	"Saya sangat menikmati cuaca hari ini."	"Saya sangat menikmati cuaca ini."	Menghapus kata acak untuk mengurangi ketergantungan pada kata tertentu.

Keempat teknik ini meningkatkan variasi kalimat, membantu model memahami cara penyampaian informasi yang berbeda. Hal ini tidak hanya meningkatkan akurasi model, tetapi juga mengurangi *overfitting*, sehingga model dapat lebih baik menggeneralisasi data uji yang belum pernah dilihat sebelumnya [15].

## 2.9. Matriks Evaluasi

Matriks evaluasi adalah sebuah tabel yang digunakan dalam penelitian untuk mengukur kinerja berbagai model pembelajaran mesin. Matriks Evaluasi sangat penting untuk memahami seberapa baik model tersebut bekerja dalam tugas klasifikasi. Beberapa metrik yang umum digunakan adalah akurasi, presisi, *recall* dan *F1 score* yang dihitung dengan menggunakan *confusion matrix* [13]. Evaluasi ini dapat dihitung dengan perhitungan berikut:

1. Akurasi adalah rasio jumlah prediksi yang benar terhadap jumlah sampel total, dengan rumus (1).

$$Akurasi = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

2. Presisi adalah rasio antara jumlah sampel positif yang teridentifikasi dengan tepat terhadap total jumlah sampel yang diprediksi sebagai positif, dengan rumus (2).

$$Presisi = \frac{(TP)}{(TP + FP)} \quad (2)$$

3. *Recall* adalah rasio antara jumlah sampel positif yang teridentifikasi dengan benar terhadap total jumlah sampel positif yang sebenarnya, dengan rumus (3).

$$Recall = \frac{(TP)}{TP + FN} \quad (3)$$

4. *F1 score* adalah rata-rata dari *recall* dan presisi, dengan rumus (4).

$$F1\ Score = \frac{2 (Presisi * Recall)}{(Presisi + Recall)} \quad (4)$$

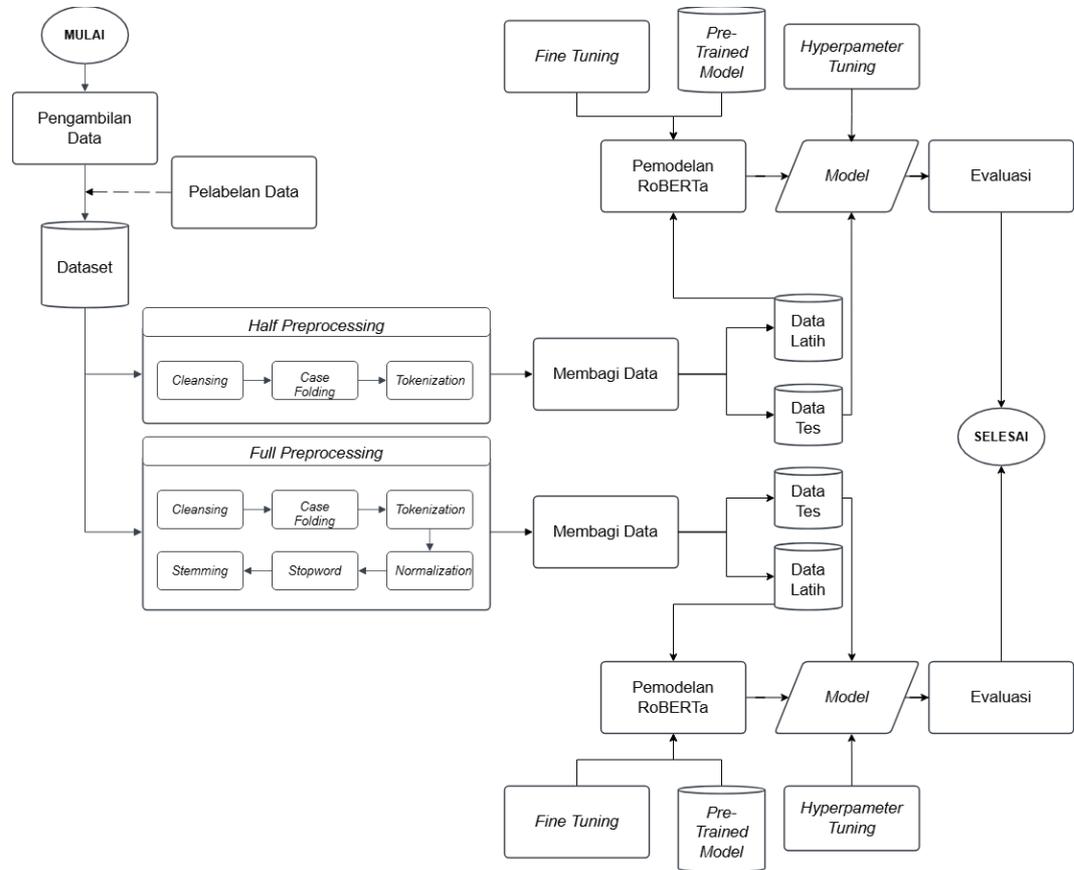
Keterangan:

- *True Positive* (TP): Kasus positif yang diprediksi dengan benar oleh model [13].
- *True Negative* (TN): Kasus negatif yang diprediksi dengan benar oleh model [13].
- *False Positive* (FP): Kasus yang diprediksi positif oleh model tetapi sebenarnya negatif [13].

- *False Negative* (FN): Kasus yang diprediksi negatif oleh model tetapi sebenarnya positif [13].

## BAB 3 PERANCANGAN SISTEM

### 3.1. Desain Perancangan Sistem



Gambar 3. 1 Alur Diagram Sistem

Alur diagram pada Gambar 3.1 menggambarkan rancangan sistem yang bertujuan untuk mengklasifikasikan kepribadian *Big Five* berdasarkan data media sosial X. Desain ini mencakup beberapa tahapan utama, yaitu pengambilan data dari media sosial, pelabelan data berdasarkan hasil kuesioner *Big Five Inventory* (BFI-44), *preprocessing data*, pemodelan menggunakan RoBERTa, dan evaluasi.

*Dataset* dibagi menjadi dua bagian untuk menguji pengaruh dua jenis *preprocessing* terhadap performa model. *Half preprocessing* mencakup tahapan dasar seperti *cleansing*, *case folding*, dan *tokenization*, yang memberi gambaran awal tentang bagaimana model bekerja dengan data yang lebih mentah. Sementara itu, *full preprocessing* melibatkan tahapan tambahan seperti *stemming*, *stopword removal*, dan *normalization*, yang bertujuan menghasilkan data lebih bersih dan

terstruktur, diharapkan dapat meningkatkan akurasi model dalam klasifikasi kepribadian.

### 3.2. Pengambilan Data

Pada penelitian ini, digunakan dua jenis data utama. Pertama, data berupa hasil kuisisioner BFI-44 yang digunakan untuk menentukan kepribadian responden berdasarkan model *Big Five Personality*. Kedua, data berupa cuitan yang dikumpulkan melalui proses *crawling* dari media sosial X, yang diperoleh dari responden yang telah menyetujui penggunaan datanya untuk keperluan penelitian ini.

### 3.3. Pelabelan Data

Proses pelabelan data dilakukan berdasarkan hasil kuisisioner *Big Five Inventory* (BFI-44) yang telah diisi oleh para responden. Kuisisioner ini memberikan penilaian terhadap lima dimensi utama kepribadian, yaitu *openness*, *conscientiousness*, *extraversion*, *agreeableness*, dan *neuroticism*. Setiap dimensi kepribadian diberi nilai tertentu berdasarkan jawaban responden, yang kemudian dikonversi menjadi label numerik. Representasi numerik ini digunakan sebagai target label dalam proses klasifikasi kepribadian, sehingga model dapat memahami dan memprediksi dimensi kepribadian dengan lebih mudah. Detail representasi label numerik dapat dilihat pada Tabel 3.1.

Tabel 3. 1 Pelabelan Data

Label	Label Numerik
<i>Openness</i>	0
<i>Conscientiousness</i>	1
<i>Extraversion</i>	2
<i>Agreeableness</i>	3
<i>Neuroticism</i>	4

### 3.4. *Preprocessing Data*

Tahap *preprocessing data* bertujuan untuk meningkatkan kualitas data teks agar siap digunakan dalam pelatihan model. Data mentah sering kali mengandung elemen-elemen tidak relevan seperti tanda baca, huruf kapital, atau kata tidak baku, yang dapat mengurangi performa model. Untuk itu, dilakukan beberapa tahapan *preprocessing*, yaitu *cleansing*, *case folding*, *tokenization*, *normalization*, *stopword*, *stemming*, serta pembagian ke dua skenario *preprocessing* yaitu *half preprocessing* dan *full preprocessing*, kedua skenario *preprocessing* ini diuji untuk melihat pengaruh kompleksitas dalam tahapan *preprocessing* terhadap akurasi model RoBERTa.

Pada *half preprocessing*, hanya tahapan dasar yang dilakukan, seperti *cleansing*, *case folding*, dan *tokenization*, yang memungkinkan model bekerja dengan data lebih mentah dan mempertahankan variasi kata. Sebaliknya, *full preprocessing* mencakup langkah-langkah tambahan seperti *normalization*, *stopword removal*, dan *stemming* untuk menghasilkan data yang lebih bersih dan terstruktur. Pengujian kedua skenario ini bertujuan untuk mengevaluasi sejauh mana kompleksitas *preprocessing* memengaruhi performa model, serta untuk mengetahui apakah data yang lebih bersih memberikan keuntungan signifikan dalam klasifikasi kepribadian. Berikut penjelasan tiap tahapan beserta contohnya:

#### 3.4.1. *Cleansing*

Proses *Cleansing* adalah proses untuk membersihkan kalimat dari karakter karakter seperti tanda baca, karakter khusus ataupun data yang tidak relevan. Contoh penerapan *cleansing* dapat dilihat pada Tabel 3.2.

Tabel 3. 2 *Cleansing*

Masukkan	Hasil
Saya mau beli Buku cerita di @tokobuku, link nya cek di <a href="http://bit.ly/bukubaru">http://bit.ly/bukubaru</a>	Saya mau beli Buku cerita di tokobuku link nya cek di

### 3.4.2. Case Folding

*Case folding* adalah proses mengubah semua huruf dalam teks menjadi huruf kecil. Tujuan dari *case folding* adalah untuk menghilangkan perbedaan yang disebabkan oleh penggunaan huruf kapital, sehingga memastikan konsistensi dalam data. Contoh penerapan *case folding* dapat dilihat pada Tabel 3.3.

Tabel 3. 3 *Case Folding*

Masukkan	Hasil
Saya mau beli Buku cerita di tokobuku link nya cek di	saya mau beli buku cerita di tokobuku link nya cek di

### 3.4.3. Tokenization

*Tokenization* adalah proses memecah teks menjadi potongan kata atau token. Contoh penerapan *tokenization* dapat dilihat pada Tabel 3.4.

Tabel 3. 4 *Tokenization*

Masukkan	Hasil
saya mau beli buku cerita di tokobuku link nya cek di	["saya", "mau", "beli", "buku", "cerita", "di", "tokobuku", "link", "nya", "cek", "di"]

### 3.4.4. Normalization

*Normalization* bertujuan mengubah kata tidak baku menjadi kata baku sesuai kaidah KBBI. Proses ini membantu menyelaraskan data sehingga lebih mudah dipahami oleh model. Contoh penerapan *normalization* dapat dilihat pada Tabel 3.5.

Tabel 3. 5 *Normalization*

Masukkan	Hasil
["saya", "mau", "beli", "buku", "cerita", "di", "tokobuku", "link", "nya", "cek", "di"]	["saya", "ingin", "membeli", "buku", "cerita", "di", "tokobuku", "tautan", "nya", "cek", "di"]

### 3.4.5. Stopword

Penghapusan *stopword* adalah proses menghapus kata-kata umum yang tidak memiliki kontribusi signifikan terhadap analisis. Contoh penerapan penghapusan *stopword* dapat dilihat pada Tabel 3.6.

Tabel 3. 6 *Stopword*

Masukkan	Hasil
["saya", "ingin", "membeli", "buku", "cerita", "di", "tokobuku", "tautan", "nya", "cek", "di"]	["ingin", "membeli", "buku", "cerita", "tokobuku", "tautan", "cek"]

### 3.4.6. Stemming

*Stemming* adalah proses mengubah kata ke bentuk dasarnya dengan menghapus imbuhan seperti awalan dan akhiran. Contoh penerapan *stemming* dapat dilihat pada Tabel 3.7.

Tabel 3. 7 *Stemming*

Masukkan	Hasil
["ingin", "membeli", "buku", "cerita", "tokobuku", "tautan", "cek"]	["ingin", "beli", "buku", "cerita", "tokobuku", "taut", "cek"]

### 3.4.7. Half Preprocessing

*Half preprocessing* digunakan untuk melihat dampak *preprocessing* yang lebih sederhana terhadap performa model. Proses ini hanya melibatkan tahapan dasar seperti *cleansing*, *case folding*, dan *tokenization*, tanpa menyentuh aspek lain seperti penghapusan *stopword* atau *stemming*. Dengan pendekatan ini, data tetap mempertahankan banyak informasi asli, yang dapat membantu model belajar dari variasi bahasa yang lebih alami. Namun, hal ini juga membuat model lebih rentan terhadap *noise*, karena tidak ada tahapan untuk mengurangi kata-kata yang kurang relevan.

Contoh hasil dari *Half Preprocessing* dapat dilihat pada Tabel 3.8, di mana data masih dalam bentuk yang lebih mentah dibandingkan dengan *Full Preprocessing*.

Tabel 3. 8 Hasil *Half Preprocessing*

Tahapan	Contoh Hasil
Cleansing	Saya mau beli Buku cerita di @tokobuku, link nya cek di <a href="http://bit.ly/bukubaru">http://bit.ly/bukubaru</a>
Case Folding	saya mau beli buku cerita di tokobuku link nya cek di
Tokenization	["saya", "mau", "beli", "buku", "cerita", "di", "tokobuku", "link", "nya", "cek", "di"]

#### 3.4.8. *Full Preprocessing*

*Full preprocessing* bertujuan untuk membersihkan dan menstrukturkan data dengan lebih teliti. Dalam metode ini, selain langkah-langkah dasar seperti *cleansing* dan *case folding*, juga dilakukan *normalization*, *stopword removal*, dan *stemming*. Tujuannya adalah untuk menghilangkan elemen yang tidak relevan, sehingga model dapat fokus pada informasi yang lebih penting.

Hasil *Full Preprocessing* memberikan data yang lebih konsisten dan bersih untuk meningkatkan akurasi model. Contoh hasil dari proses ini dapat dilihat pada Tabel 3.9.

Tabel 3. 9 Hasil *Full Preprocessing*

Tahapan	Contoh Hasil
Cleansing	Saya mau beli Buku cerita di @tokobuku, link nya cek di <a href="http://bit.ly/bukubaru">http://bit.ly/bukubaru</a>
Case Folding	saya mau beli buku cerita di tokobuku link nya cek di
Tokenization	["saya", "mau", "beli", "buku", "cerita", "di", "tokobuku", "link", "nya", "cek", "di"]
Normalization	["saya", "ingin", "membeli", "buku", "cerita", "di", "tokobuku", "tautan", "nya", "cek", "di"]
Stopword Removal	["ingin", "membeli", "buku", "cerita", "tokobuku", "tautan", "cek"]
Stemming	["ingin", "beli", "buku", "cerita", "tokobuku", "taut", "cek"]

### 3.5. *Hyperparameter Tuning*

Pada tahap *Hyperparameter Tuning*, dilakukan pencarian parameter yang optimal untuk model RoBERTa menggunakan metode *Random Search*. Tujuan dari tahap ini adalah untuk menemukan kombinasi parameter terbaik yang dapat meningkatkan kinerja model pada tugas klasifikasi kepribadian *Big Five*. Beberapa parameter yang diuji meliputi *batch size*, *learning rate*, dan *weight decay*.

Pencarian *hyperparameter* dilakukan menggunakan *dataset full preprocessing* dan *half preprocessing*. Hasil yang diperoleh akan diterapkan pada model yang akan dilatih. Proses ini berlangsung selama 100 iterasi, dengan *F1-score* sebagai metrik utama untuk menentukan performa terbaik. Dengan cara ini, parameter yang diperoleh diharapkan dapat memberikan performa optimal terlepas dari pembagian data yang digunakan dalam pelatihan dan pengujian model.

### 3.6. Pemodelan Roberta

Pemodelan RoBERTa dalam penelitian ini digunakan untuk mengklasifikasikan kepribadian *Big Five* berdasarkan data cuitan media sosial X. RoBERTa dilatih dengan teknik *dynamic masking*, yang menyembunyikan kata-kata secara acak untuk memperkaya pemahaman konteks antar kata, sehingga menghasilkan representasi teks yang lebih kuat dibandingkan dengan pendekatan *static masking* pada BERT [4].

Model dilatih dengan membagi data menjadi dua bagian: data pelatihan dan data pengujian. Meskipun kode hanya memisahkan data menjadi *train/test*, *validation loss* dihitung menggunakan data pengujian. Ini dilakukan dengan mengonfigurasi *eval\_dataset* pada *Trainer*, yang memungkinkan model untuk dievaluasi pada setiap epoch menggunakan data yang tidak terlihat selama pelatihan pada Gambar 3.2 terdapat code untuk inisiasi trainer dengan model, data, dan parameter pelatihan. Dengan cara ini, data pengujian berfungsi sebagai *dataset* validasi untuk menghitung *validation loss*, serta metrik seperti akurasi, *F1-score*, presisi, dan *recall*.

```
trainer = Trainer(  
    model=model,  
    args=training_args, # Parameter pelatihan  
    train_dataset=train_dataset, # Dataset pelatihan  
    eval_dataset=test_dataset, # Dataset pengujian  
    tokenizer=tokenizer, # Tokenizer untuk mengubah teks menjadi token  
    compute_metrics=compute_metrics, # Fungsi untuk menghitung metrik evaluasi  
    callbacks=[early_stopping_callback] # Menambahkan callback untuk early stopping  
)
```

Gambar 3. 2 Code Inisiasi Trainer Model RoBERTa

Model ini di *fine-tune* setelah tahap *hyperparameter tuning*, di mana parameter seperti *learning rate*, *batch size*, dan *weight decay* diuji untuk menemukan kombinasi yang optimal. Pengaturan terbaik dari hasil tuning tersebut digunakan untuk *fine-tuning* model. Penyesuaian ini menunjukkan bahwa penggunaan *batch size* lebih besar dan *learning rate* yang lebih tinggi dapat meningkatkan performa model. Setelah *fine-tuning*, RoBERTa menghasilkan representasi teks yang digunakan untuk memprediksi dimensi kepribadian pengguna [4].

### 3.7. Pendekatan Model

Dalam penelitian ini, empat pendekatan model diuji untuk melihat seberapa besar dampaknya terhadap performa dalam klasifikasi kepribadian *Big Five* menggunakan model RoBERTa. Keempat pendekatan ini dipilih untuk mengeksplorasi berbagai teknik yang dapat meningkatkan kemampuan model dalam menganalisis data teks dari media sosial. Setiap pendekatan diterapkan pada *dataset* yang telah diproses dengan dua jenis *preprocessing*, yaitu *half preprocessing* dan *full preprocessing*. Pendekatan-pendekatan ini diuji menggunakan dua jenis pembagian data, yaitu pembagian *data training* dan *data test* yang terdiri dari 80/20 dan 70/20.

#### 3.7.1. *Baseline*

Pendekatan *Baseline* adalah model dasar yang dilatih dengan data yang telah melalui tahap *preprocessing*, tanpa adanya penambahan teknik seperti augmentasi data atau penyeimbangan kelas. Pendekatan ini memberikan gambaran awal tentang bagaimana model bekerja dengan data yang telah diproses dan berfungsi sebagai referensi untuk membandingkan hasil dengan pendekatan lainnya yang lebih kompleks.

#### 3.7.2. *Data Augmentation*

Pada pendekatan *Data Augmentation*, teknik untuk memperbanyak variasi dalam *dataset* diterapkan. *Easy Data Augmentation* (EDA) digunakan untuk memperkaya data dengan cara mengganti kata dengan sinonim, menyisipkan kata acak, menukar posisi kata, atau menghapus kata secara acak. Teknik ini bertujuan untuk membantu model belajar dari lebih banyak variasi teks, yang diharapkan dapat mengurangi *overfitting* dan meningkatkan kemampuan model dalam mengenali pola yang lebih beragam [15].

Penambahan variasi dengan EDA membantu model mengenali pola yang lebih beragam, sangat berguna untuk *dataset* kecil atau tidak seimbang. Teknik ini mengurangi *overfitting* dan meningkatkan kemampuan model untuk menggeneralisasi data baru, menghasilkan performa yang lebih stabil dan akurat pada data uji [15].

Misalnya, dengan Synonym Replacement (SR), kata-kata dalam cuitan diganti dengan sinonim, sehingga model bisa memahami konsep lebih luas. Random Swap (RS), di sisi lain, mengubah urutan kata untuk memperkenalkan fleksibilitas dalam memahami struktur kalimat. Teknik ini diharapkan meningkatkan metrik evaluasi seperti F1-score, akurasi, precision, dan recall, terutama pada dataset yang terbatas [15].

### **3.7.3. *Random Oversampling (ROS)***

Pendekatan *Random Oversampling (ROS)* diterapkan untuk mengatasi ketidakseimbangan kelas dalam *dataset*. ROS bekerja dengan menggandakan data dari kelas minoritas untuk menyeimbangkan distribusi kelas. Dengan demikian, model tidak akan lebih fokus pada kelas mayoritas dan dapat mempelajari pola dari kelas minoritas dengan lebih baik [14].

### **3.7.4. *ROS + Data Augmentation***

Pendekatan *Data Augmentation + ROS* menggabungkan kedua teknik tersebut untuk menghasilkan *dataset* yang lebih seimbang dan bervariasi. Dengan menggabungkan augmentasi data dan *random oversampling*, diharapkan model dapat memanfaatkan kelebihan dari kedua teknik ini untuk menghasilkan klasifikasi yang lebih baik, mengurangi bias terhadap kelas mayoritas, dan meningkatkan keragaman data yang digunakan dalam pelatihan.

## **3.8. Evaluasi**

Selanjutnya, dilakukan tahap evaluasi untuk mengukur kinerja model. Evaluasi ini dilakukan dengan menggunakan metode *Confusion Matrix* yang sangat berguna untuk menguji akurasi model. *Confusion Matrix* memiliki empat karakteristik utama, yaitu *True Positive (TP)*, *True Negative (TN)*, *False Positive (FP)*, dan *False Negative (FN)*. Kinerja model diukur menggunakan beberapa metrik, antara lain akurasi, presisi, *recall*, dan *F1-Score*, yang dihitung berdasarkan nilai TP, TN, FP, dan FN [13]. Mengacu pada rumusan yang dijelaskan di sub bab 2.9, berikut adalah contoh perhitungan nilai evaluasi yang tertera pada Tabel 3.10.

Tabel 3. 10 Contoh nilai evaluasi

<b>Aktual</b>	<b>Prediksi</b>	
	<i>True</i>	<i>False</i>
<i>Positive</i>	97	43
<i>Negative</i>	102	52

$$Akurasi = \frac{97+102}{97+43+102+52} = \frac{199}{294} = 0,676$$

$$Presisi = \frac{97}{97+43} = \frac{87}{140} = 0,693$$

$$Recall = \frac{97}{97+52} = \frac{97}{149} = 0,651$$

$$F1\ score = \frac{2(0,693 * 0,651)}{0,693+0,651} = \frac{2 * (0,451)}{1,344} = \frac{0,902}{1,344} = 0,671$$

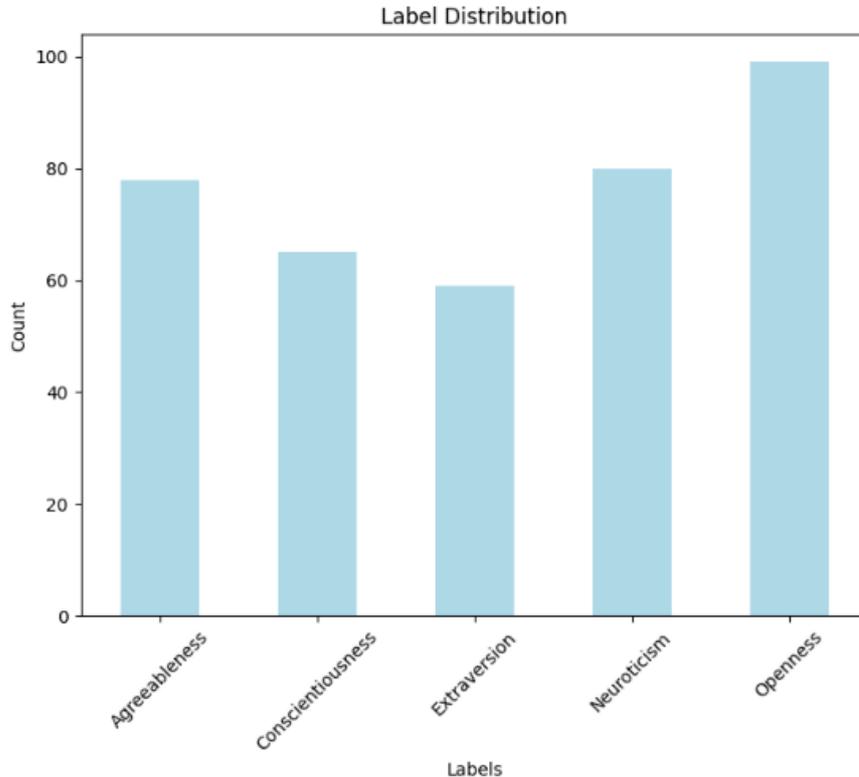
Setelah melakukan contoh perhitungan pada tahapan evaluasi, dapat disimpulkan bahwa nilai akurasi pada model yang diuji yaitu sebesar 0,676, presisi sebesar 0,693, *recall* sebesar 0,651, dan *F1 score* sebesar 0,671.

## BAB 4 HASIL PERCOBAAN DAN ANALISIS

### 4.1. Dataset

Pada penelitian ini, *dataset* yang digunakan untuk mengklasifikasikan kepribadian *Big Five* diambil dari platform media sosial X. Data ini merupakan gabungan antara hasil survei BFI-44, yang berisi username akun media sosial beserta kecenderungan kepribadian *Big Five Personality* dari responden. Setiap akun diwakili oleh lima dimensi kepribadian: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, dan *Neuroticism*. Cuitan dari masing-masing akun kemudian digabungkan dan digunakan sebagai data teks untuk melatih model.

Distribusi label dalam *dataset* ini cukup tidak seimbang, dengan jumlah data yang lebih banyak pada beberapa label. Dapat dilihat pada label *Openness* memiliki jumlah data yang lebih tinggi, sementara label *Extraversion* memiliki jumlah data yang lebih rendah. Hal ini dapat dilihat pada Gambar 4.1, yang menunjukkan distribusi jumlah data untuk masing-masing label *Big Five*.



Gambar 4. 1 Distribusi Label *Dataset*

Tabel 4.1 memperlihatkan distribusi label pada *dataset* yang digunakan, dengan rincian jumlah data sebagai berikut:

- *Openness*: 99
- *Neuroticism*: 80
- *Agreeableness*: 78
- *Conscientiousness*: 65
- *Extraversion*: 59

Kedua *dataset* yang digunakan, yaitu *Full Preprocessing* dan *Half Preprocessing*, berisi cuitan yang diambil dari 381 pengguna, dengan total data sebanyak 381 baris. Perbedaan antara keduanya terletak pada tahapan *preprocessing* yang diterapkan. *Half Preprocessing* mencakup tahapan *cleansing*, *case folding*, dan *tokenization*, sedangkan *Full Preprocessing* mencakup tahapan *cleansing*, *case folding*, *tokenization*, *normalization*, *stopword removal*, dan *stemming*.

Ketidakseimbangan label ini menjadi tantangan dalam percobaan ini, yang kemudian diatasi dengan teknik seperti *Data Augmentation* dan *Random Oversampling* (ROS). Teknik-teknik ini digunakan untuk menyeimbangkan distribusi kelas dan meningkatkan akurasi model.

#### 4.2. *Hyperparameter Tuning*

Pada penelitian ini, *Hyperparameter Tuning* dilakukan untuk mengoptimalkan kinerja model RoBERTa dalam mengklasifikasikan kepribadian *Big Five*. Proses ini bertujuan untuk menemukan kombinasi *hyperparameter* terbaik yang dapat meningkatkan performa model. Teknik *Random Search* digunakan untuk mengeksplorasi berbagai kombinasi parameter yang mungkin menghasilkan kinerja terbaik pada model. *Hyperparameter* yang diuji mencakup beberapa parameter penting, yaitu *Batch Size*, *Learning Rate*, dan *Weight Decay*, yang berpengaruh signifikan terhadap hasil pelatihan model.

Pada proses *tuning*, *Hyperparameter* yang diuji meliputi:

- *Batch Size*: Jumlah sampel yang diproses dalam satu iterasi pelatihan.

- *Learning Rate*: Tingkat pembelajaran yang mengontrol seberapa besar perubahan parameter model pada setiap iterasi.
- *Weight Decay*: Faktor regularisasi yang digunakan untuk mencegah *overfitting* dengan memberikan penalti terhadap bobot model yang besar.

Nilai-nilai untuk *hyperparameter* tersebut dipilih berdasarkan distribusi yang telah ditentukan sebelumnya:

- *Batch size*: [8, 16, 32]
- *Learning rate*: [1e-5, 5e-5]
- *Weight decay*: [0.01, 0.05]

Proses *Hyperparameter Tuning* dilakukan dengan *Random Search*, yang memilih kombinasi parameter secara acak dari rentang nilai yang ditentukan. Proses ini dilakukan dalam 100 iterasi, di mana untuk setiap iterasi, nilai *batch size*, *learning rate*, dan *weight decay* dipilih secara acak. Setiap kombinasi *hyperparameter* yang dipilih diterapkan pada model dan kemudian dievaluasi menggunakan data uji.

Selama proses pelatihan, *optimizer* AdamW digunakan untuk mengoptimalkan model, dengan *CrossEntropyLoss* sebagai fungsi *loss* yang digunakan untuk tugas klasifikasi multi kelas. Model dilatih pada data pelatihan, dan setelah setiap iterasi, evaluasi dilakukan pada data uji dengan menghitung metrik akurasi, presisi, *recall*, dan *F1 Score*. Hasil evaluasi ini memberikan gambaran mengenai kinerja model pada kombinasi *hyperparameter* yang diuji.

Hasil dari proses *Hyperparameter Tuning* menunjukkan kombinasi parameter terbaik untuk masing-masing jenis *preprocessing* yaitu *Full Preprocessing* dan *Half Preprocessing*, yang memberikan *F1 Score* tertinggi:

- a. *Full Preprocessing*:
  - *Batch size*: 32
  - *Learning rate*: 2.1187e-05
  - *Weight decay*: 0.0438

Pada *Full Preprocessing*, model menunjukkan performa terbaik dengan *F1 Score* sebesar 0.4604, yang menunjukkan kombinasi parameter ini mampu

memberikan keseimbangan yang baik antara presisi dan *recall* dalam mengklasifikasikan kepribadian *Big Five*.

b. *Half Preprocessing*:

- *Batch size*: 32
- *Learning rate*: 4.5174e-05
- *Weight decay*: 0.0468

Pada *Half Preprocessing*, meskipun menggunakan *batch size* yang sama, nilai *F1 Score* yang diperoleh lebih rendah, yaitu 0.3826, dibandingkan dengan *Full Preprocessing*. Hal ini menunjukkan bahwa *preprocessing* yang lebih kompleks memberikan kontribusi yang lebih besar dalam meningkatkan performa model.

Hasil dari proses *Hyperparameter Tuning* menunjukkan kombinasi parameter terbaik yang akan digunakan untuk melatih model pada seluruh *dataset* dan mengukur kinerjanya.

### 4.3. Skenario Percobaan

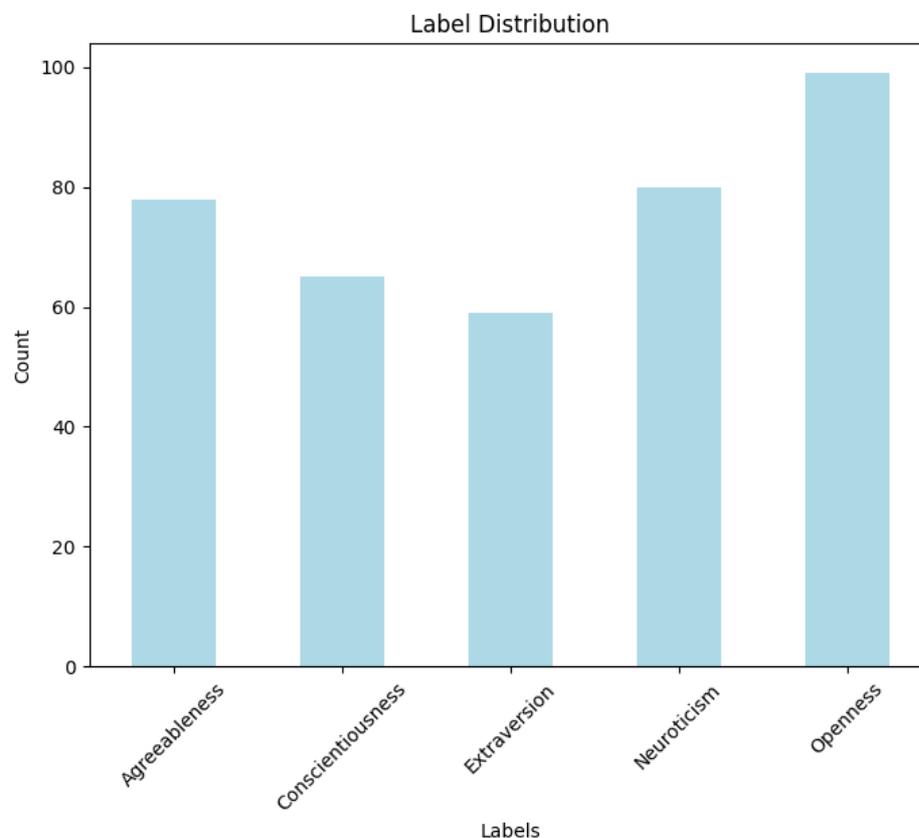
Pada penelitian ini, dilakukan serangkaian percobaan untuk mengklasifikasikan kepribadian *Big Five* berdasarkan data teks dari media sosial X. Percobaan ini bertujuan untuk mengevaluasi pendekatan terbaik dalam meningkatkan performa model. Langkah pertama yang dilakukan adalah *preprocessing data*, yang terbagi menjadi dua skenario, yaitu *Full Preprocessing* dan *Half Preprocessing*. Perbedaan utama antara keduanya terletak pada tingkat pembersihan data, di mana *Full Preprocessing* mencakup lebih banyak tahapan seperti normalisasi, penghapusan *stopword*, dan *stemming*. Skenario ini dirancang untuk melihat bagaimana kompleksitas *preprocessing* memengaruhi hasil klasifikasi.

Setelah melalui tahap *preprocessing*, data dibagi ke dalam dua skenario pembagian, yaitu 80/20 (80% untuk pelatihan dan 20% untuk pengujian) serta 70/30 (70% untuk pelatihan dan 30% untuk pengujian). Tujuan dari perbedaan proporsi ini adalah untuk mengukur sejauh mana jumlah data latih memengaruhi performa model. Selanjutnya, setiap kombinasi *preprocessing* dan pembagian data

diuji dengan empat pendekatan berbeda: *Baseline*, *Random Oversampling (ROS)*, *Data Augmentation*, dan kombinasi *Data Augmentation + ROS*. Keempat pendekatan ini diterapkan untuk mengamati dampak teknik augmentasi dan penyeimbangan data terhadap hasil klasifikasi. Seluruh skenario percobaan dilakukan menggunakan *hyperparameter* yang telah disesuaikan melalui proses *tuning*, sehingga model dapat bekerja secara optimal di setiap pengujian. Dengan kombinasi *preprocessing*, pembagian data, dan pendekatan yang diterapkan, terdapat total 16 skenario percobaan yang digunakan untuk mengevaluasi model dalam tugas klasifikasi kepribadian *Big Five*.

#### 4.3.1. Skenario 1 : *Baseline*

Pada skenario ini, model diuji tanpa teknik tambahan seperti augmentasi data atau penyeimbangan kelas. Skenario ini memberikan gambaran awal tentang bagaimana performa model tanpa modifikasi data lebih lanjut. Hasil dari *Full Preprocessing* dan *Half Preprocessing* juga dibandingkan untuk melihat dampak kompleksitas preprocessing terhadap hasil yang diperoleh.



Gambar 4. 2 Distribusi Label *Baseline*

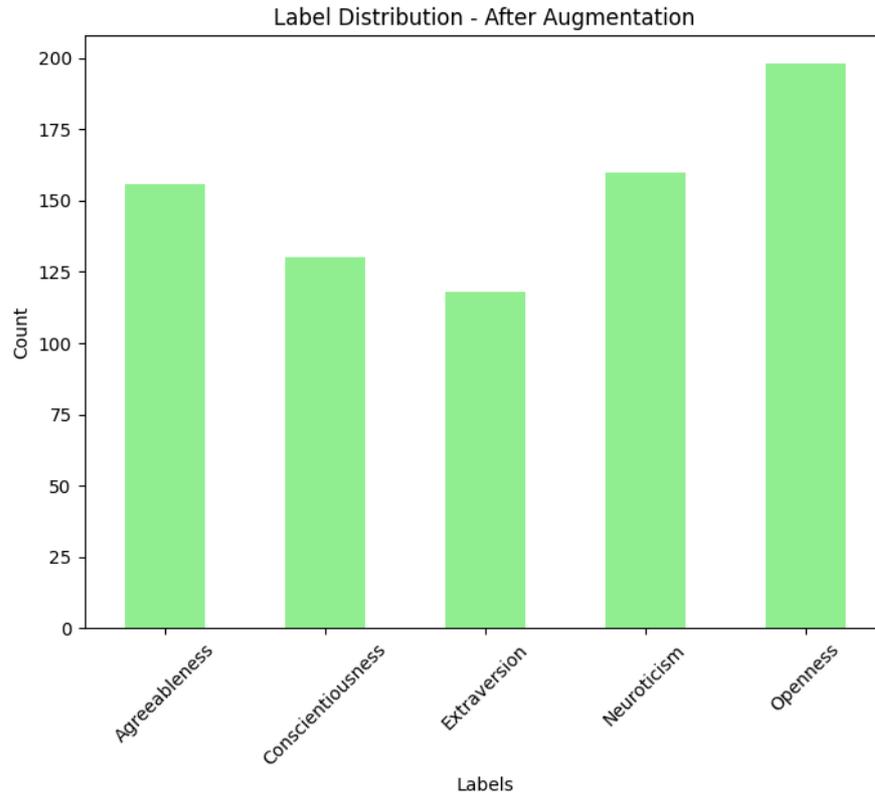
Jumlah data pada skenario *Baseline* dapat dilihat pada Gambar 4.2, yang menunjukkan distribusi label *Big Five* yang tidak seimbang. Berikut adalah jumlah data untuk tiap label pada skenario *Baseline*:

- *Openness*: 99
- *Neuroticism*: 80
- *Agreeableness*: 78
- *Conscientiousness*: 65
- *Extraversion*: 59

Ketidakseimbangan data ini dapat memengaruhi hasil model, terutama dalam hal akurasi prediksi pada label yang memiliki jumlah data lebih sedikit.

#### **4.3.2. Skenario 2 : *Data Augmentation***

Pada skenario ini, teknik *Data Augmentation* diterapkan dengan tujuan untuk memperkaya *dataset*. Teknik *Easy Data Augmentation* (EDA) digunakan untuk memperkenalkan variasi dalam data melalui beberapa metode, seperti *synonym replacement*, *random insertion*, *random swap*, dan *random deletion*. Dengan teknik ini, model akan lebih mampu mengenali pola yang lebih beragam dan mengurangi *overfitting*. Teknik ini diuji pada *dataset* yang telah melalui *Full Preprocessing* dan *Half Preprocessing* untuk melihat perbedaan kinerja.



Gambar 4. 3 Distribusi Label Setelah *Data Augmentation*

Jumlah data pada skenario *Data Augmentation* dapat dilihat pada Gambar 4.3, yang menunjukkan perubahan pada distribusi label setelah augmentasi data. Berikut adalah jumlah data untuk tiap label setelah teknik augmentasi diterapkan:

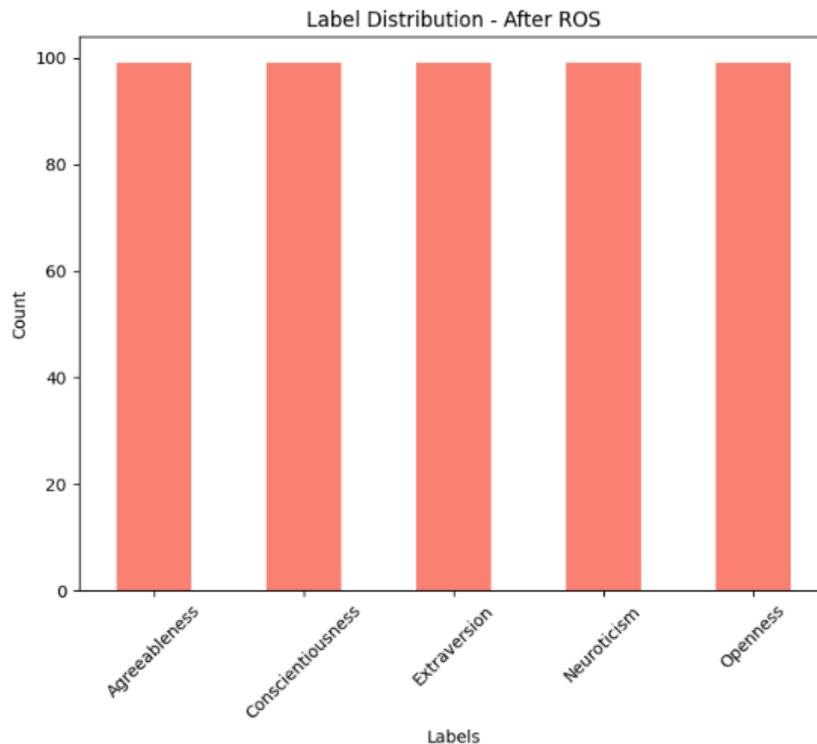
- *Openness*: 198
- *Neuroticism*: 160
- *Agreeableness*: 156
- *Conscientiousness*: 130
- *Extraversion*: 118

Distribusi label menjadi lebih merata, meskipun label *Extraversion* masih memiliki jumlah yang lebih sedikit dibandingkan label lainnya.

#### 4.3.3. Skenario 3 : *Random Oversampling (ROS)*

Pada skenario ini, teknik *Random Oversampling (ROS)* digunakan untuk menyeimbangkan distribusi kelas dalam *dataset*. ROS menggandakan data dari

kelas minoritas, sehingga memastikan model tidak terlalu bias terhadap kelas mayoritas. Teknik ini diharapkan membuat model lebih baik dalam mempelajari pola dari kelas minoritas. Seperti pada skenario lainnya, ROS diuji pada data yang melalui *Full Preprocessing* dan *Half Preprocessing* untuk mengevaluasi pengaruh teknik ini terhadap model.



Gambar 4. 4 Distribusi Label Setelah *Random Oversampling* (ROS)

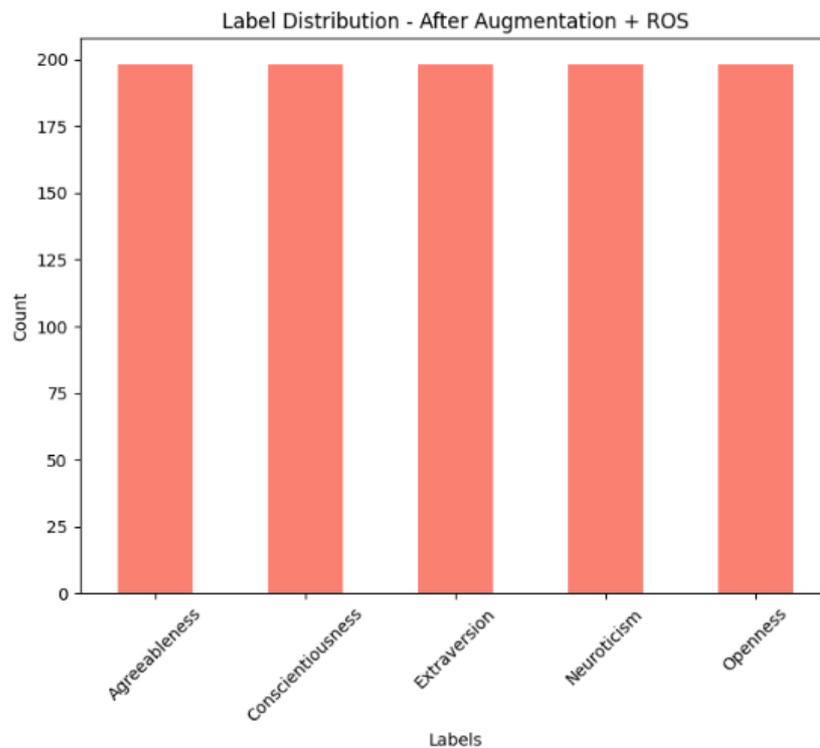
Distribusi label setelah ROS dapat dilihat pada Gambar 4.4, yang menunjukkan distribusi label setelah menggunakan teknik ROS. Jumlah data untuk tiap label pada skenario ROS adalah:

- *Openness*: 99
- *Neuroticism*: 99
- *Agreeableness*: 99
- *Conscientiousness*: 99
- *Extraversion*: 99

Dengan teknik ROS, distribusi label menjadi seimbang, memberikan model kesempatan yang sama untuk mempelajari pola pada setiap kelas.

#### 4.3.4. Skenario 4 : *Data Augmentation* + ROS

Pada skenario ini, *Data Augmentation* digabungkan dengan *Random Oversampling* (ROS) untuk menghasilkan *dataset* yang lebih seimbang dan bervariasi. Gabungan kedua teknik ini bertujuan untuk memperkaya *dataset* dengan variasi baru dan memastikan distribusi data antar kelas menjadi lebih merata. Teknik ini juga diuji pada data yang melalui *Full Preprocessing* dan *Half Preprocessing* untuk mengevaluasi pengaruh gabungan kedua teknik ini terhadap model.



Gambar 4. 5 Distribusi Label Setelah *Data Augmentation* + ROS

Distribusi label setelah *Data Augmentation* + ROS dapat dilihat pada Gambar 4.5, yang menunjukkan distribusi label yang sangat seimbang setelah kedua teknik diterapkan. Berikut adalah jumlah data untuk tiap label:

- *Openness*: 198
- *Neuroticism*: 198
- *Agreeableness*: 198
- *Conscientiousness*: 198
- *Extraversion*: 198

Dengan penggabungan teknik ini, setiap label memiliki jumlah data yang sama, yang memastikan bahwa model belajar dari distribusi kelas yang lebih seimbang.

#### 4.4. Hasil Percobaan

Pada bagian ini, hasil dari setiap percobaan berdasarkan rasio pembagian data 80/20 dan 70/30 serta dua jenis preprocessing yaitu *Full Preprocessing* dan *Half Preprocessing* akan dipaparkan. Hasil percobaan mencakup beberapa metrik evaluasi, seperti Akurasi, *F1 Score*, Presisi, dan *Recall*.

##### 4.4.1. Hasil Percobaan 1 :*Baseline*

Pada percobaan pertama ini, model diuji tanpa menggunakan teknik tambahan seperti *Data Augmentation* atau *Random Oversampling* (ROS). Model dilatih dengan dua jenis *preprocessing*, yaitu *Full Preprocessing* dan *Half Preprocessing*, dan diuji pada dua pembagian data yang berbeda, yaitu 80/20 dan 70/30. Evaluasi model dilakukan menggunakan beberapa metrik utama, seperti Akurasi, *F1 Score*, Presisi, dan *Recall*. Hasil dari percobaan ini dapat dilihat pada Tabel 4.1.

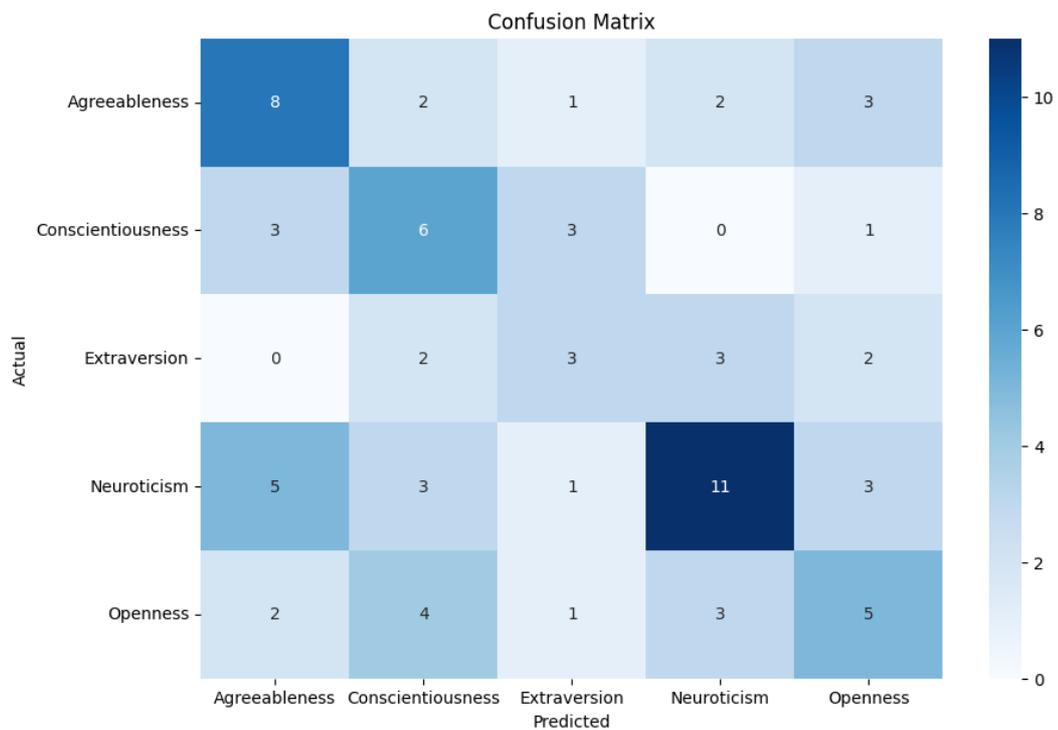
Berdasarkan hasil evaluasi, penerapan *Full Preprocessing* memberikan performa yang lebih baik dibandingkan dengan *Half Preprocessing*. Pada pembagian data 80/20, model dengan *Full Preprocessing* mencapai Akurasi 42.86% dan *F1 Score* 42.99%, sebagaimana terlihat pada Gambar 4.6, yang menunjukkan distribusi prediksi yang lebih merata pada sebagian besar dimensi kepribadian. Pada skenario pembagian data 70/30, *Full Preprocessing* menghasilkan Akurasi 30.43% dan *F1 Score* 27.87%, yang digambarkan pada Gambar 4.7, yang menunjukkan distribusi prediksi model yang lebih tersebar.

Di sisi lain, *Half Preprocessing* menunjukkan penurunan performa pada kedua pembagian data. Pada 80/20, model dengan *Half Preprocessing* mencatat Akurasi 35.06% dan *F1 Score* 35.19%, sebagaimana terlihat pada Gambar 4.8, yang menunjukkan kesalahan prediksi yang lebih tinggi pada beberapa dimensi. Pada pembagian data 70/30, performa model lebih menurun lagi dengan Akurasi 24.34% dan *F1 Score* 15.44%, yang terlihat pada Gambar 4.9, yang menggambarkan

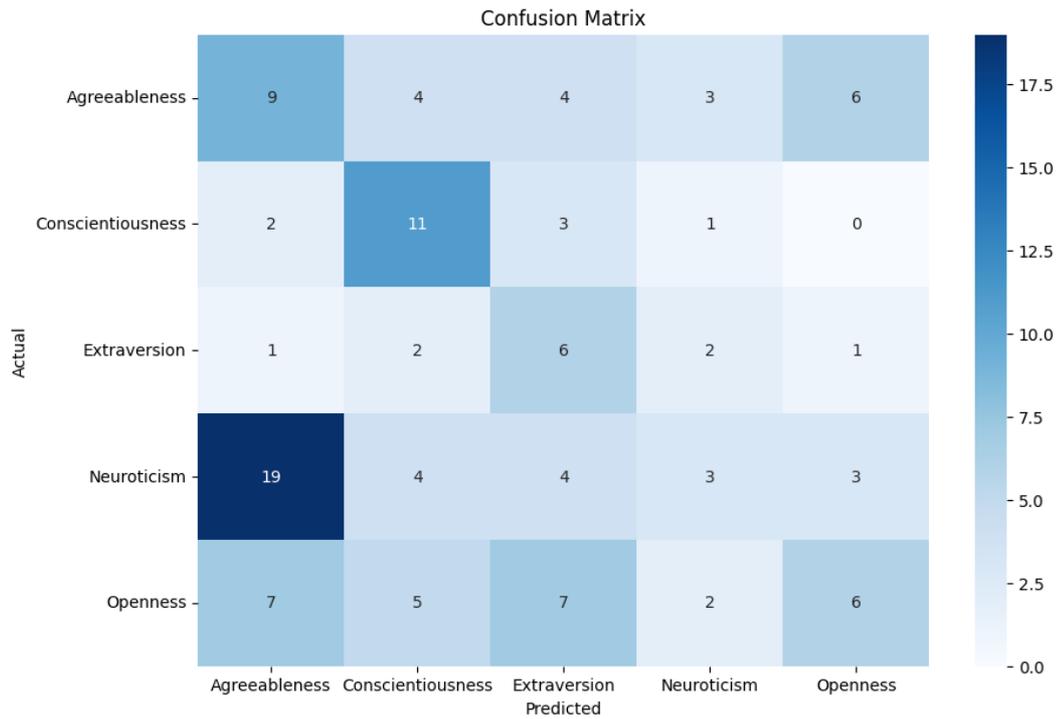
kesalahan prediksi yang signifikan pada dimensi *Openness*, *Extraversion*, dan *Neuroticism*.

Tabel 4. 1 Hasil Percobaan Skenario 1 *Baseline*

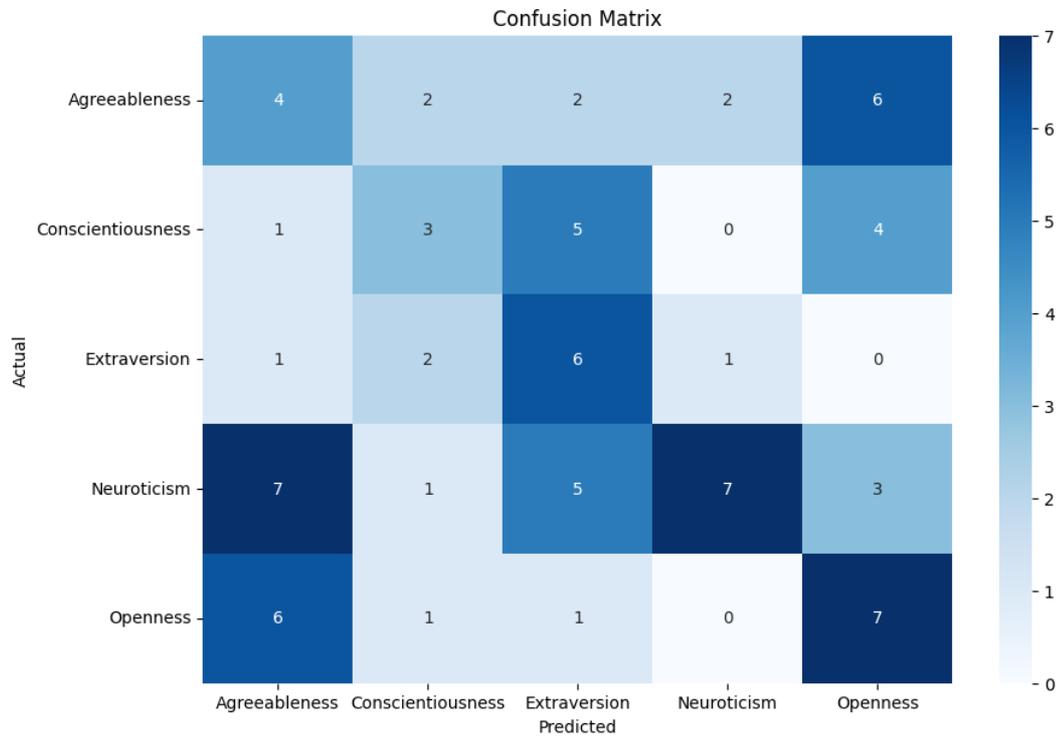
Jenis Preprocessing	Rasio Pembagian Data	Akurasi(%)	<i>F1 Score</i> (%)	Presisi(%)	<i>Recall</i> (%)
<i>Full Preprocessing</i>	80/20	42.86	42.99	43.77	42.86
<i>Preprocessing</i>	70/30	30.43	27.87	28.12	30.0
<i>Half Preprocessing</i>	80/20	35.06	35.19	41.83	35.06
<i>Preprocessing</i>	70/30	24.34	15.44	11.32	24.34



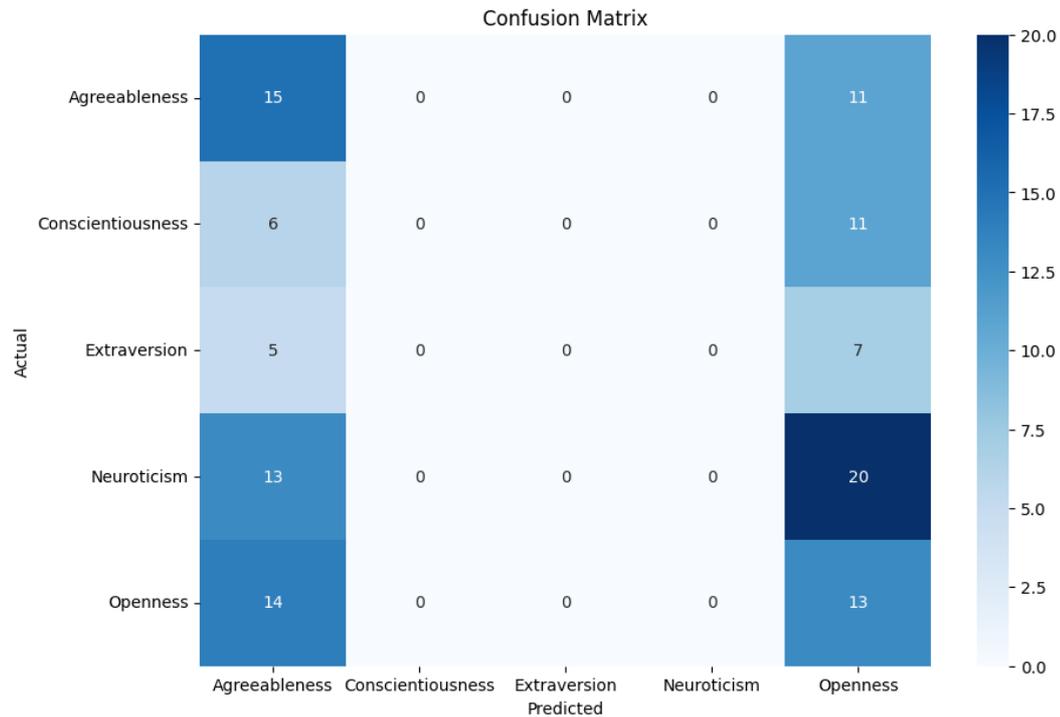
Gambar 4. 6 *Confusion Matrix* untuk *Baseline* dengan *Full Preprocessing* (80/20)



Gambar 4. 7 *Confusion Matrix* untuk *Baseline* dengan *Full Preprocessing* (70/30)



Gambar 4. 8 *Confusion Matrix* untuk *Baseline* dengan *Half Preprocessing* (80/20)



Gambar 4. 9 *Confusion Matrix* untuk *Baseline* dengan *Half Preprocessing* (70/30)

*Confusion Matrix* memberikan wawasan lebih mendalam mengenai distribusi prediksi model pada setiap dimensi kepribadian. Sebagai contoh, pada Gambar 4.6, model dengan *Full Preprocessing* pada pembagian data 80/20 berhasil memprediksi dimensi *Agreeableness* dan *Neuroticism* dengan cukup baik, meskipun masih terdapat beberapa kesalahan prediksi pada dimensi lain, seperti *Extraversion* dan *Openness*. Sebaliknya, Pada Gambar 4.9, hasil prediksi model dengan *Half Preprocessing* dan pembagian data 70/30 menunjukkan distribusi yang tidak merata pada beberapa dimensi kepribadian, seperti *Agreeableness*, *Conscientiousness*, dan *Openness*. Misalnya, model sering keliru memprediksi *Agreeableness* sebagai *Openness*. *Conscientiousness* dan *Extraversion* juga menunjukkan kesalahan yang cukup tinggi. Ini terjadi karena *Half Preprocessing* yang lebih sederhana tidak cukup untuk mengatasi data yang tidak seimbang, di mana beberapa dimensi kepribadian kurang terwakili dalam dataset.

Hasil percobaan *Baseline* ini memperlihatkan bahwa *preprocessing* yang lebih lengkap, seperti *Full Preprocessing*, memiliki dampak positif terhadap performa model. Namun, distribusi data yang tidak merata masih menjadi tantangan utama, terutama untuk dimensi seperti *Openness* dan *Extraversion*, yang sering kali

diprediksi dengan tingkat kesalahan yang tinggi. Untuk mengatasi tantangan ini, percobaan berikutnya akan menguji teknik tambahan seperti *Data Augmentation* dan *Random Oversampling* (ROS) untuk meningkatkan distribusi data dan performa model secara keseluruhan.

#### **4.4.2. Hasil Percobaan 2 : *Data Augmentation***

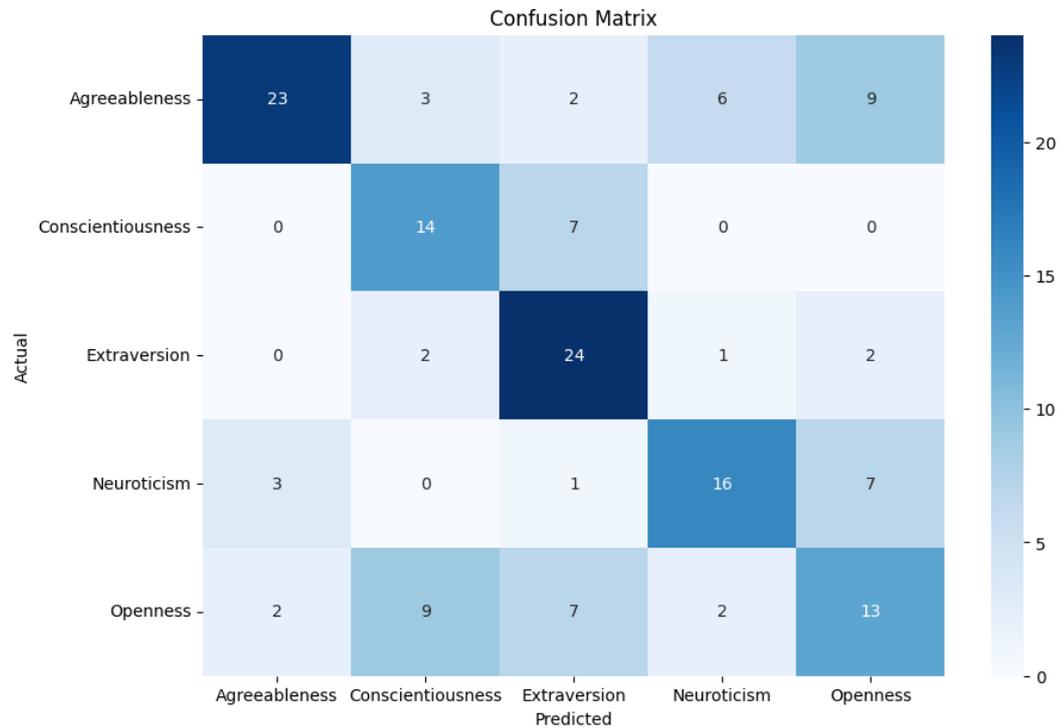
Pada percobaan kedua ini, model diberi pelatihan menggunakan teknik *Data Augmentation* untuk memperluas variasi data yang ada. Teknik ini diterapkan pada dua jenis *preprocessing*, yaitu *Full Preprocessing* dan *Half Preprocessing*, dan diuji pada dua pembagian data yang berbeda, yaitu 80/20 dan 70/30. Evaluasi model dilakukan dengan mengukur beberapa metrik seperti Akurasi, *F1 Score*, Presisi, dan *Recall*. Hasil dari percobaan ini dapat dilihat pada Tabel 4.2.

Berdasarkan hasil evaluasi, penerapan *Data Augmentation* dengan *Full Preprocessing* memberikan hasil terbaik, terutama pada pembagian data 80/20. Model dengan *Full Preprocessing* pada skenario ini mencapai Akurasi 58.82% dan *F1 Score* 58.67%, sebagaimana terlihat pada Gambar 4.10, yang menggambarkan distribusi prediksi model yang lebih merata pada dimensi kepribadian yang lebih sulit diprediksi. Pada pembagian data 70/30, *Full Preprocessing* menghasilkan Akurasi 52.40% dan *F1 Score* 52.41%, yang dapat dilihat pada Gambar 4.11, yang menunjukkan distribusi prediksi yang lebih merata.

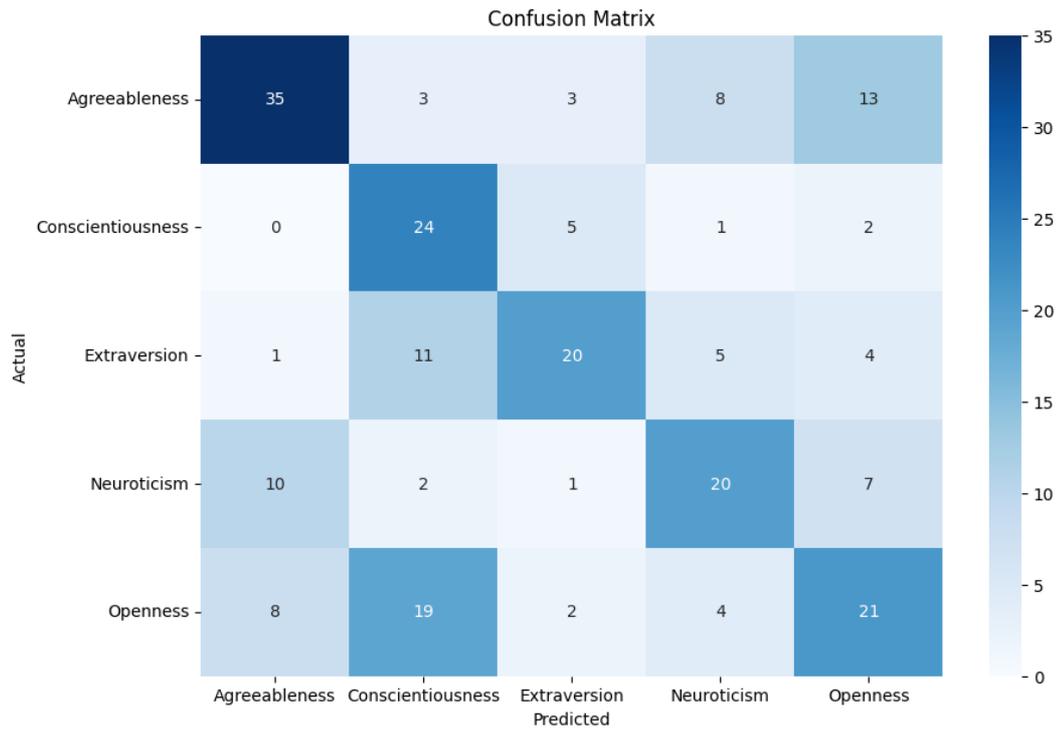
Di sisi lain, penerapan *Half Preprocessing* memberikan sedikit peningkatan dibandingkan dengan percobaan *Baseline*, namun performanya tetap lebih rendah dibandingkan *Full Preprocessing*, terutama pada pembagian data 70/30. Pada skenario ini, *Half Preprocessing* hanya mencapai Akurasi 34.06% dan *F1 Score* 29.09%, yang terlihat pada Gambar 4.13, yang menunjukkan distribusi prediksi yang kurang seimbang. Sementara itu, pada pembagian data 80/20, *Half Preprocessing* hanya mencapai Akurasi 44.44% dan *F1 Score* 35.06%, sebagaimana terlihat pada Gambar 4.12, yang menunjukkan distribusi prediksi yang lebih bervariasi.

Tabel 4. 2 Hasil Percobaan Skenario 2 *Data Augmentation*

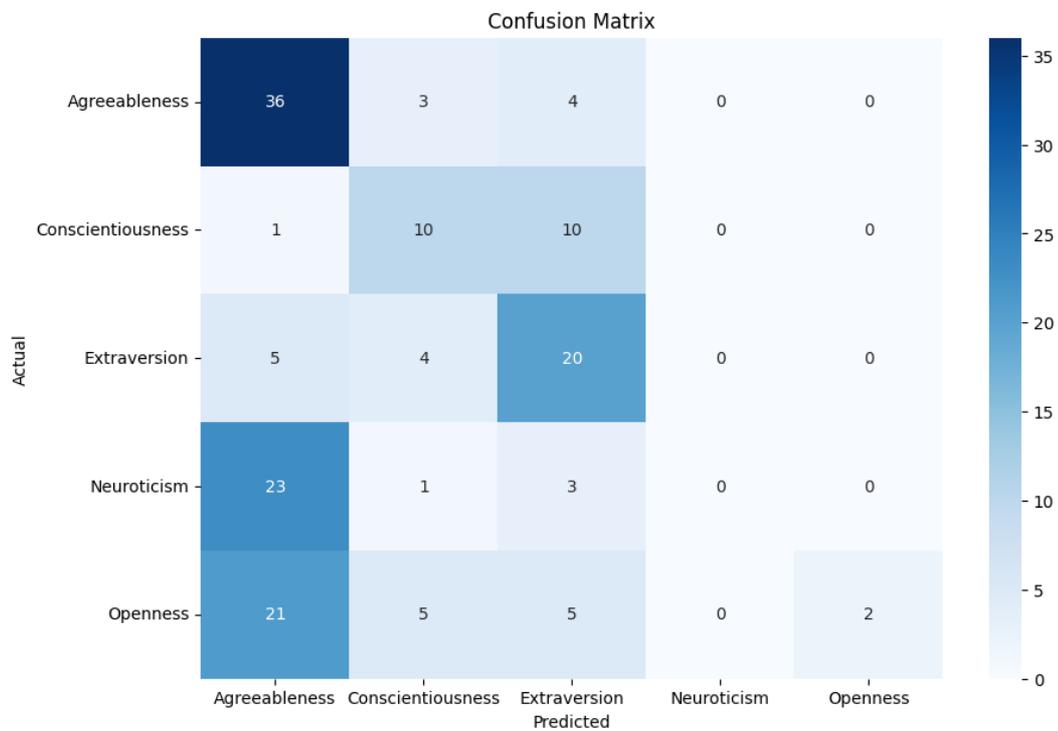
Jenis <i>Preprocessing</i>	Rasio Pembagian Data	Akurasi(%)	<i>F1 Score</i> (%)	Presisi(%)	<i>Recall</i> (%)
<i>Full</i>	80/20	58.82	58.67	61.38	58.82
<i>Preprocessing</i>	70/30	52.40	52.41	54.51	52.40
<i>Half</i>	80/20	44.44	35.06	48.32	44.44
<i>Preprocessing</i>	70/30	34.06	29.09	47.03	34.06



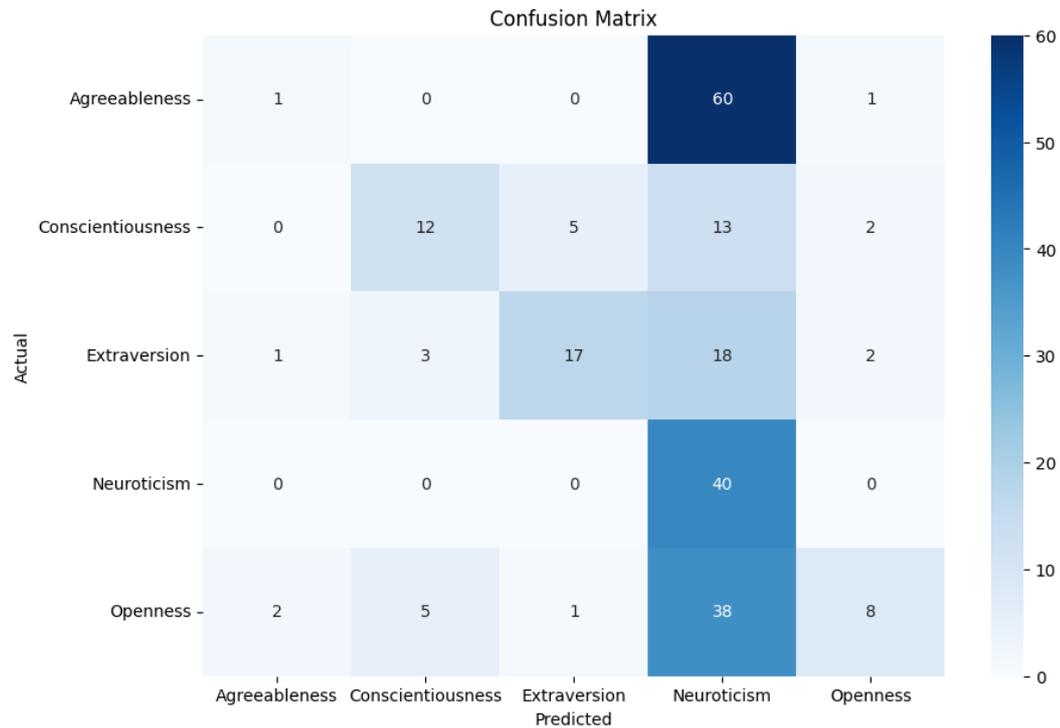
Gambar 4. 10 *Confusion Matrix* untuk *Data Augmentation* dengan *Full Preprocessing* (80/20)



Gambar 4. 11 *Confusion Matrix* untuk *Data Augmentation* dengan *Full Preprocessing* (70/30)



Gambar 4. 12 *Confusion Matrix* untuk *Data Augmentation* dengan *Half Preprocessing* (80/20)



Gambar 4. 13 *Confusion Matrix* untuk *Data Augmentation* dengan *Half Preprocessing* (70/30)

*Confusion Matrix* memberikan wawasan mendalam mengenai distribusi prediksi model pada setiap dimensi kepribadian. Sebagai contoh, pada Gambar 4.10, model dengan *Full Preprocessing* pada pembagian data 80/20 dapat memprediksi dimensi *Agreeableness* dan *Extraversion* dengan cukup akurat, meskipun masih terdapat kesalahan prediksi pada dimensi *Conscientiousness* dan *Openness*. Sebaliknya, pada Gambar 4.13, model dengan *Half Preprocessing* pada pembagian data 70/30 menunjukkan distribusi prediksi yang jauh lebih tidak merata, dengan tingkat kesalahan yang tinggi pada dimensi *Agreeableness*, *Conscientiousness*, dan *Openness*.

Penerapan *Data Augmentation* pada *Full Preprocessing* berhasil memperbaiki distribusi prediksi pada dimensi yang lebih sulit seperti *Extraversion*. Teknik ini memberikan variasi lebih banyak bagi model untuk belajar mengenali pola yang sebelumnya tidak terjangkau oleh model dalam skenario *Baseline*. Meskipun demikian, *Half Preprocessing* masih menunjukkan keterbatasannya dalam menghasilkan prediksi yang lebih baik, yang mengindikasikan perlunya metode *preprocessing* yang lebih kompleks dan lebih mendalam. Berdasarkan hasil

ini, percobaan berikutnya akan menguji teknik *Random Oversampling*, yang dapat lebih menyeimbangkan distribusi data dan memperbaiki kinerja model.

#### 4.4.3. Hasil Percobaan 3 : *Random Oversampling* (ROS)

Pada percobaan ketiga ini, model dilatih menggunakan teknik *Random Oversampling* (ROS), ROS ini memperbaiki distribusi kelas yang tidak seimbang, terutama dengan menambah kelas minoritas, seperti *Extraversion*, yang sebelumnya memiliki jumlah data yang sangat rendah. Dengan teknik ini, model tidak lagi cenderung bias terhadap kelas mayoritas (*Openness*), dan hasilnya menunjukkan perbaikan signifikan pada distribusi prediksi dan penurunan kesalahan pada dimensi *Extraversion*. Teknik ini diterapkan pada dua jenis *preprocessing*, yaitu *Full Preprocessing* dan *Half Preprocessing*, serta diuji pada dua skenario pembagian data yang berbeda, yaitu 80/20 dan 70/30. Evaluasi model dilakukan dengan mengukur beberapa metrik seperti Akurasi, *F1 Score*, Presisi, dan *Recall*. Hasil percobaan ini dapat dilihat pada Tabel 4.3.

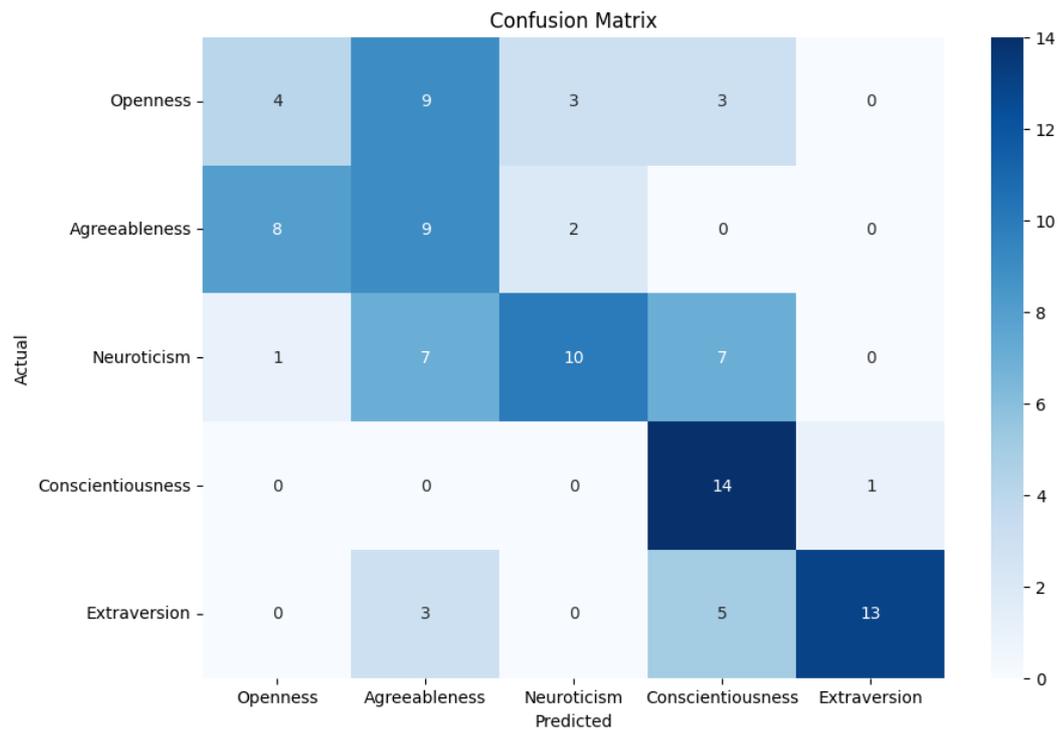
Berdasarkan hasil evaluasi, penerapan ROS dengan *Full Preprocessing* memberikan performa terbaik, terutama pada pembagian data 80/20. Model dengan *Full Preprocessing* pada skenario ini mencapai Akurasi 50.51% dan *F1 Score* 50.17%, sebagaimana terlihat pada Gambar 4.14, yang menggambarkan distribusi prediksi model terhadap dimensi *Big Five*. Namun, performa model pada pembagian data 70/30 mengalami penurunan signifikan untuk kedua jenis *preprocessing*. Pada skenario ini, *Full Preprocessing* menghasilkan Akurasi 40.94% dan *F1 Score* 37.95%, seperti yang dapat dilihat pada Gambar 4.15, yang menunjukkan bahwa distribusi data yang lebih kecil mempengaruhi kinerja model secara keseluruhan.

Sementara itu, *Half Preprocessing* juga menunjukkan penurunan performa pada kedua pembagian data. Pada 80/20, model dengan *Half Preprocessing* mencatat Akurasi 46.46% dan *F1 Score* 44.74%, yang sedikit lebih rendah dibandingkan dengan *Full Preprocessing*, seperti yang terlihat pada Gambar 4.16. Pada pembagian data 70/30, performa model menurun lebih tajam, dengan Akurasi 36.91% dan *F1 Score* 34.98%, yang terlihat pada Gambar 4.17, menunjukkan

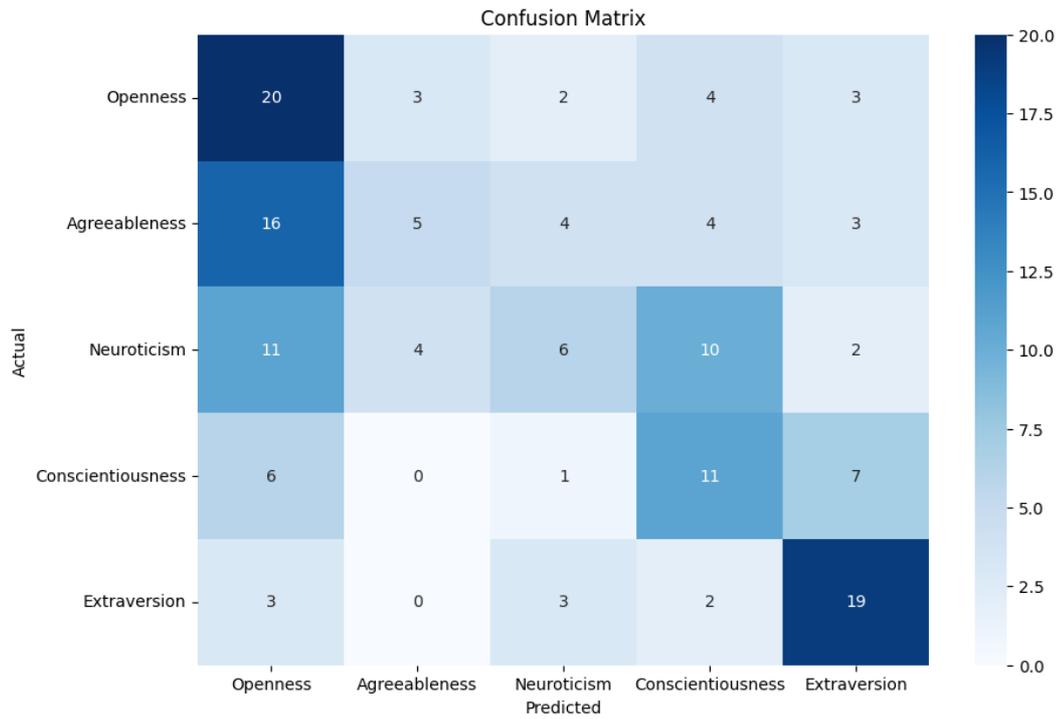
kesulitan model dalam mengenali pola yang relevan dengan jumlah data pelatihan yang terbatas.

Tabel 4. 3 Hasil Percobaan Skenario 3 *Random Oversampling* (ROS)

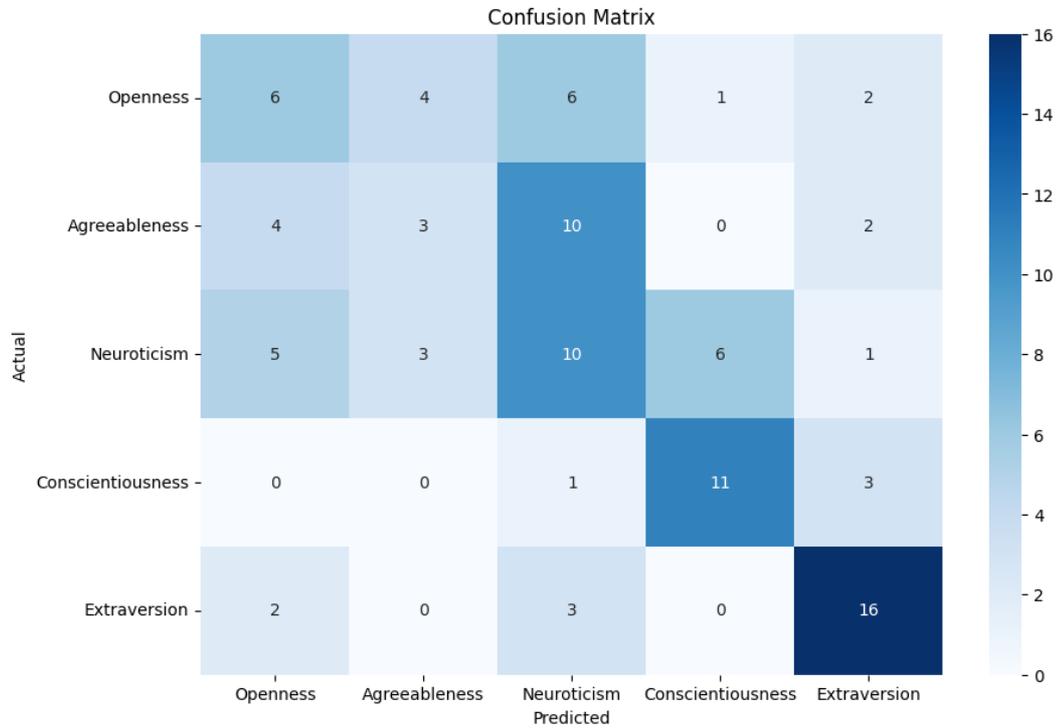
Jenis <i>Preprocessing</i>	Rasio Pembagian Data	Akurasi(%)	<i>F1 Score</i> (%)	Presisi(%)	<i>Recall</i> (%)
<i>Full</i>	80/20	50.51	50.17	55.92	50.51
<i>Preprocessing</i>	70/30	40.94	37.95	38.5	38.0
<i>Half</i>	80/20	46.46	44.74	44.35	46.46
<i>Preprocessing</i>	70/30	36.91	34.98	40.13	36.5



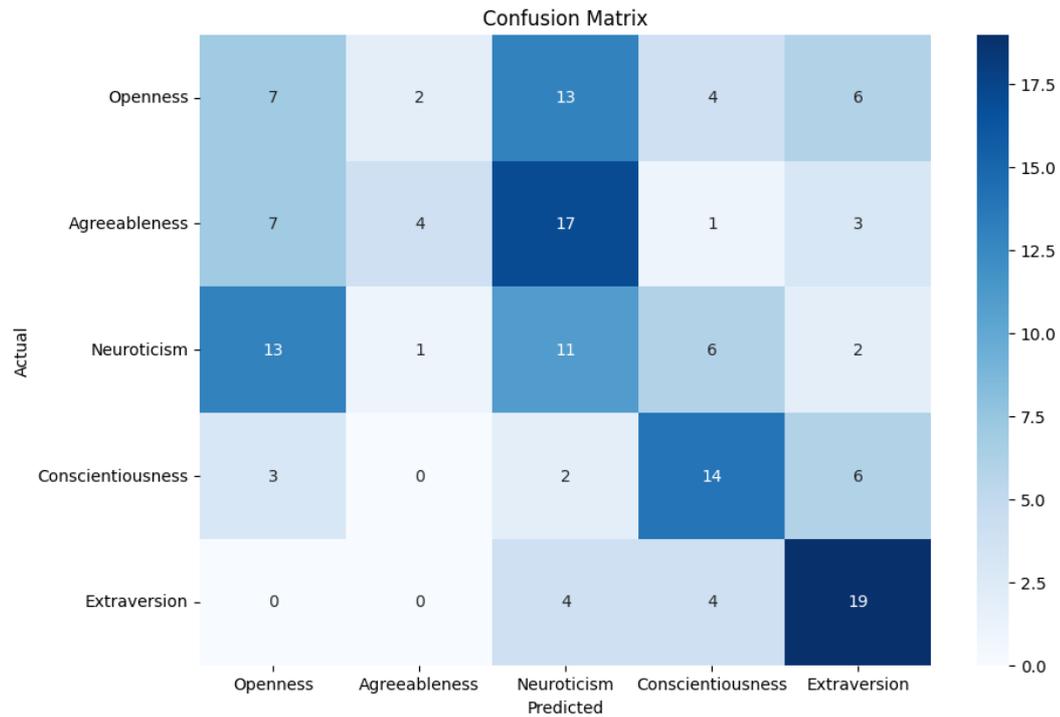
Gambar 4. 14 *Confusion Matrix* untuk *Random Oversampling* dengan *Full Preprocessing* (80/20)



Gambar 4. 15 *Confusion Matrix* untuk *Random Oversampling* dengan *Full Preprocessing* (70/30)



Gambar 4. 16 *Confusion Matrix* untuk *Random Oversampling* dengan *Half Preprocessing* (80/20)



Gambar 4. 17 *Confusion Matrix* untuk *Random Oversampling* dengan *Half Preprocessing* (70/30)

*Confusion Matrix* memberikan wawasan lebih mendalam mengenai bagaimana model memprediksi setiap dimensi kepribadian. Sebagai contoh, pada Gambar 4.14, model dengan *Full Preprocessing* pada pembagian data 80/20 dapat memprediksi dimensi *Agreeableness* dan *Neuroticism* dengan cukup akurat, meskipun terdapat kesalahan prediksi pada beberapa dimensi lain seperti *Openness* dan *Extraversion*. Sebaliknya, pada Gambar 4.17, model dengan *Half Preprocessing* pada pembagian data 70/30 menunjukkan distribusi prediksi yang lebih tidak merata, dengan kesalahan prediksi yang lebih tinggi pada dimensi *Agreeableness*, *Neuroticism*, dan *Extraversion*.

Secara keseluruhan, percobaan ini menunjukkan bahwa teknik *Random Oversampling* (ROS) berhasil meningkatkan performa model, terutama pada dimensi yang lebih sulit diprediksi pada percobaan *Baseline*. Meskipun demikian, *Half Preprocessing* masih menunjukkan keterbatasannya dalam menghasilkan prediksi yang lebih baik, yang mengindikasikan perlunya penerapan teknik *preprocessing* yang lebih kompleks dan mendalam. Berdasarkan hasil ini, percobaan berikutnya akan menggabungkan *Data Augmentation* dengan *Random*

*Oversampling* (ROS) untuk lebih menyeimbangkan distribusi data dan memperbaiki kinerja model.

#### **4.4.4. Hasil Percobaan 4 : *Data Augmentation* + ROS**

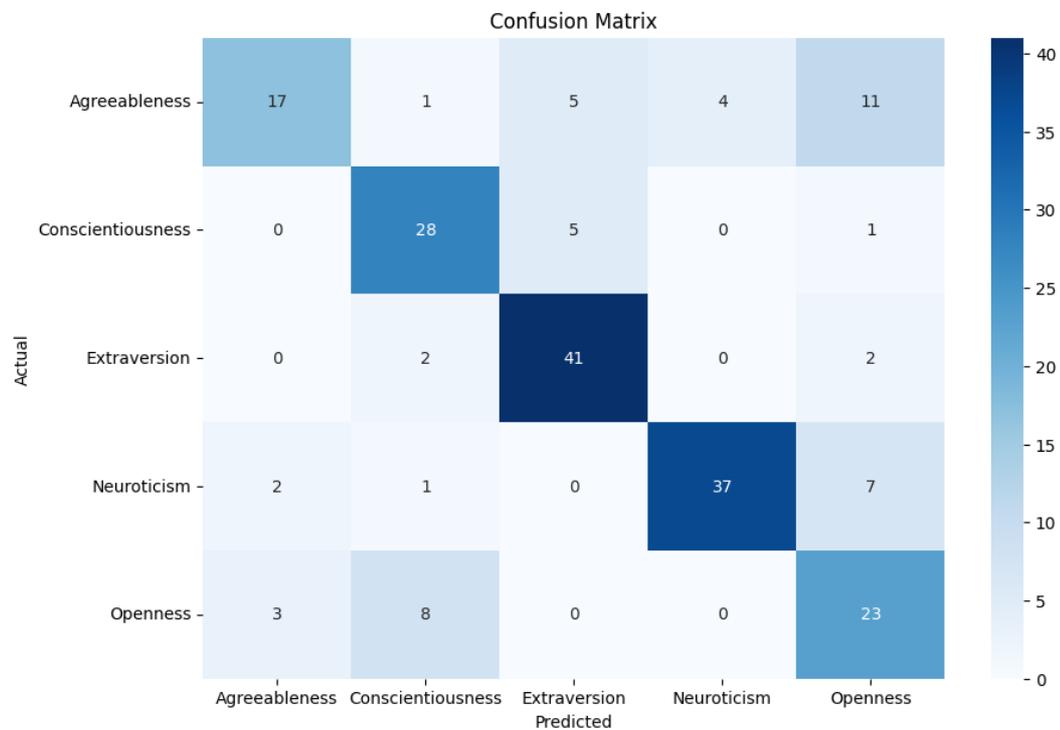
Pada percobaan keempat ini, model dilatih dengan gabungan teknik *Data Augmentation* dengan *Random Oversampling* (ROS) untuk menciptakan *dataset* yang lebih seimbang dan bervariasi. Gabungan teknik ini diterapkan pada dua jenis *preprocessing*, yaitu *Full Preprocessing* dan *Half Preprocessing*, dan diuji dengan dua pembagian data yang berbeda: 80/20 dan 70/30. Evaluasi model dilakukan berdasarkan beberapa metrik termasuk Akurasi, *F1 Score*, Presisi, dan *Recall*. Hasil dari percobaan ini dapat dilihat pada Tabel 4.4.

Berdasarkan hasil evaluasi, penerapan kombinasi *Data Augmentation* + ROS dengan *Full Preprocessing* menghasilkan performa terbaik dibandingkan semua percobaan sebelumnya. Pada skenario pembagian data 80/20, model dengan *Full Preprocessing* mencatat Akurasi tertinggi sebesar 73.73% dan *F1 Score* 73.37%, seperti yang terlihat pada Gambar 4.18. Hasil ini menunjukkan distribusi prediksi yang lebih merata pada sebagian besar dimensi kepribadian, dengan peningkatan yang signifikan pada dimensi-dimensi yang sebelumnya sulit diprediksi. Pada skenario pembagian data 70/30, model dengan *Full Preprocessing* tetap menunjukkan performa yang sangat baik, dengan Akurasi 71.38% dan *F1 Score* 71.35%, sebagaimana digambarkan pada Gambar 4.19.

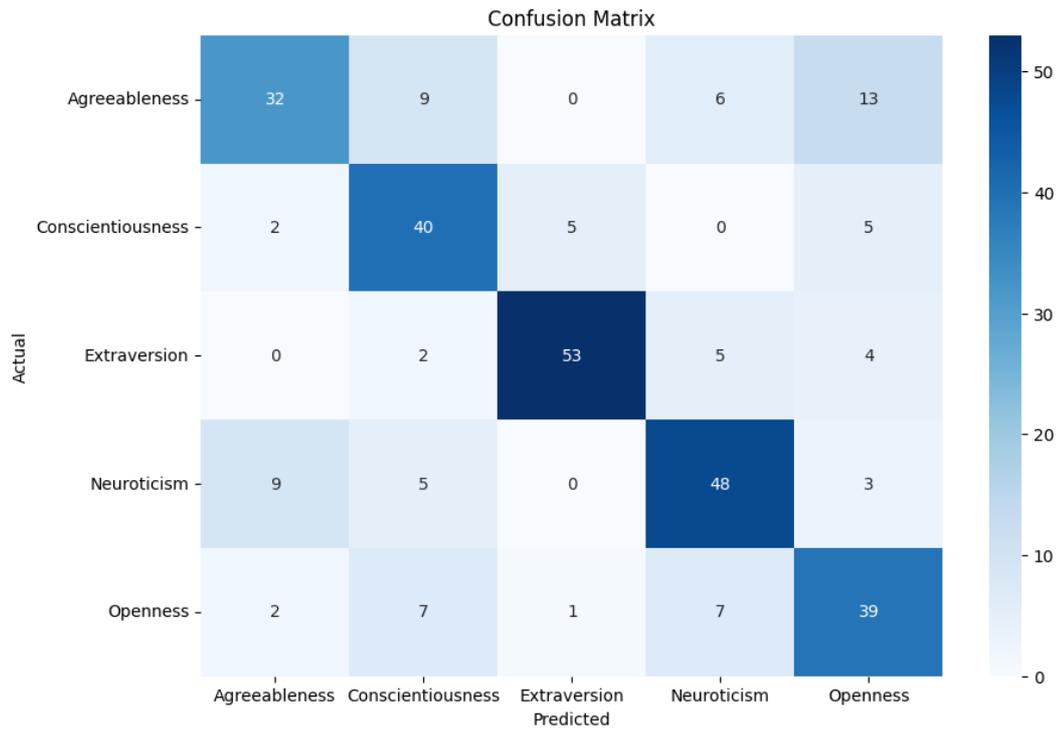
Di sisi lain, *Half Preprocessing* juga menunjukkan peningkatan performa dibandingkan dengan percobaan sebelumnya. Pada pembagian data 80/20, model dengan *Half Preprocessing* mencapai Akurasi 64.14% dan *F1 Score* 63.22%, seperti yang terlihat pada Gambar 4.20. Walaupun hasil ini lebih rendah dibandingkan *Full Preprocessing*, namun masih menunjukkan perbaikan dibandingkan dengan percobaan tanpa teknik tambahan. Pada pembagian data 70/30, *Half Preprocessing* hanya mencapai Akurasi 56.22% dan *F1 Score* 56.11%, yang terlihat pada Gambar 4.21, yang menunjukkan bahwa meskipun ada peningkatan, model dengan *Half Preprocessing* masih terbatas dalam mengenali pola yang lebih kompleks.

Tabel 4. 4 Hasil Percobaan Skenario 4 *Data Augmentation* + ROS

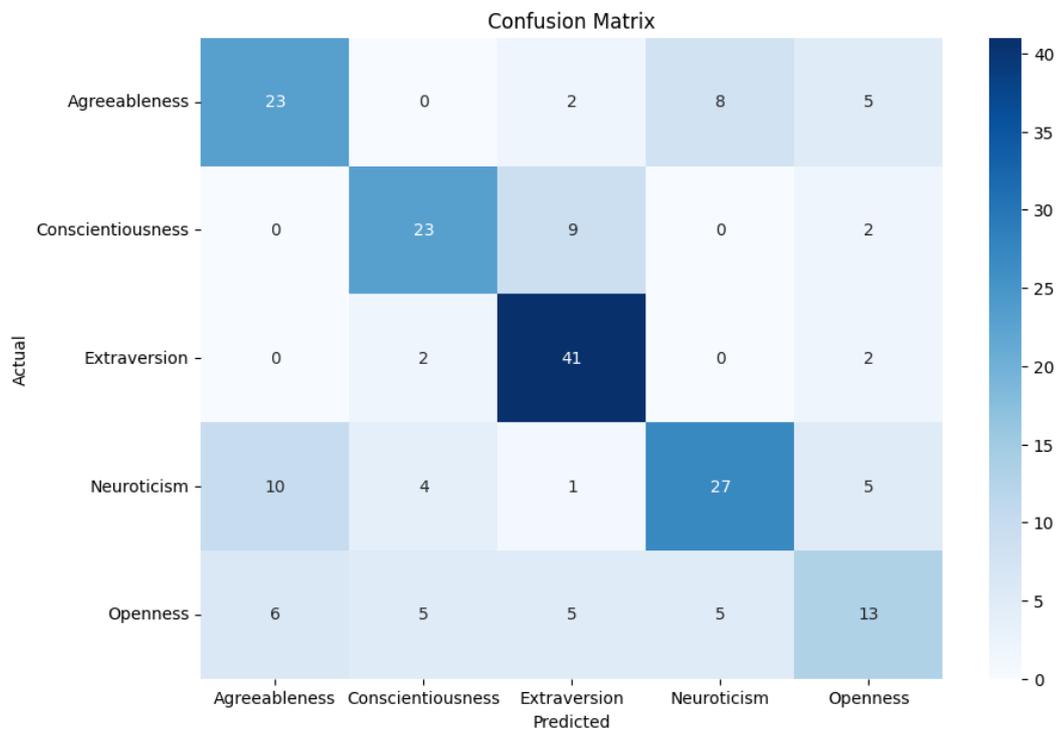
Jenis <i>Preprocessing</i>	Rasio Pembagian Data	Akurasi(%)	<i>F1 Score</i> (%)	Presisi(%)	<i>Recall</i> (%)
<i>Full</i>	80/20	73.73	73.37	75.51	73.73
<i>Preprocessing</i>	70/30	71.38	71.35	72.24	71.38
<i>Half</i>	80/20	64.14	63.22	63.29	64.14
<i>Preprocessing</i>	70/30	56.22	56.11	57.56	56.22



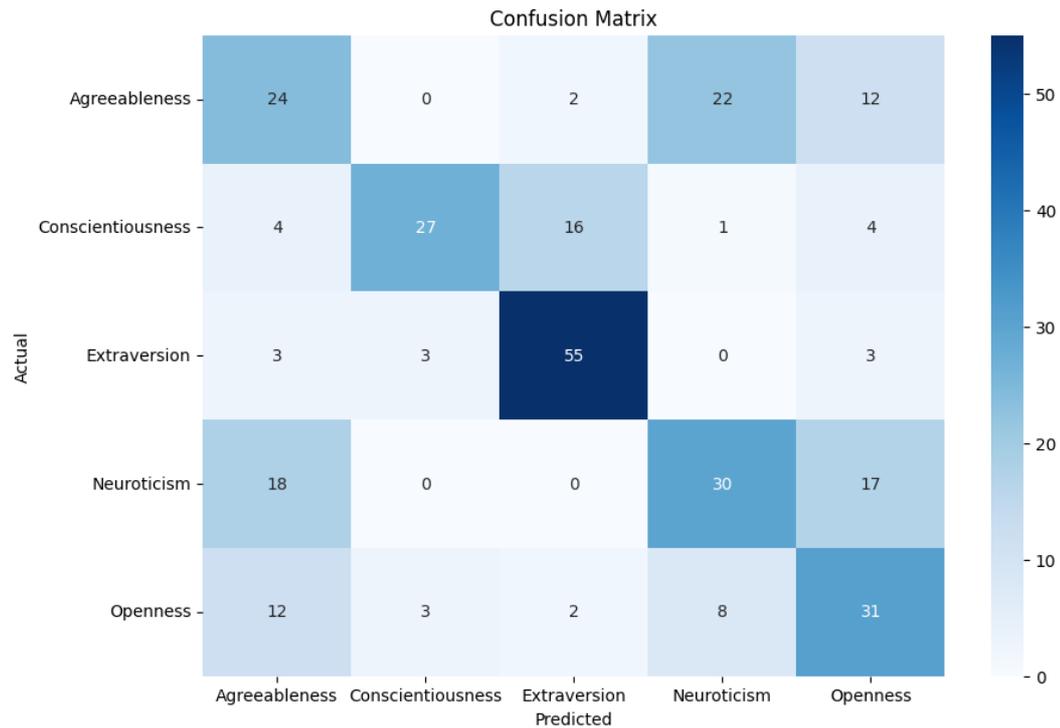
Gambar 4. 18 *Confusion Matrix* untuk *Data Augmentation* + ROS dengan *Full Preprocessing* (80/20)



Gambar 4. 19 *Confusion Matrix* untuk *Data Augmentation* + ROS dengan *Full Preprocessing* (70/30)



Gambar 4. 20 *Confusion Matrix* untuk *Data Augmentation* + ROS dengan *Half Preprocessing* (80/20)



Gambar 4. 21 *Confusion Matrix* untuk *Data Augmentation* + ROS dengan *Half Preprocessing* (70/30)

*Confusion Matrix* memberikan gambaran yang lebih mendalam mengenai distribusi prediksi model terhadap setiap dimensi kepribadian. Sebagai contoh, pada Gambar 4.18, model dengan *Full Preprocessing* pada pembagian data 80/20 berhasil memprediksi dimensi *Extraversion* dengan cukup baik, meskipun terdapat beberapa kesalahan prediksi pada dimensi *Agreeableness* dan *Openness*. Sementara itu, pada Gambar 4.21, model dengan *Half Preprocessing* pada pembagian data 70/30 menunjukkan distribusi prediksi yang lebih tidak merata, dengan kesalahan prediksi yang lebih tinggi pada dimensi *Openness*, *Neuroticism*, dan *Agreeableness*. Dimensi *Openness* khususnya menunjukkan jumlah kesalahan prediksi yang tinggi.

Secara keseluruhan, percobaan ini menunjukkan bahwa kombinasi *Data Augmentation* dan *Random Oversampling* (ROS) berhasil meningkatkan performa model secara signifikan, terutama pada dimensi-dimensi yang lebih sulit diprediksi dalam percobaan *Baseline*. Meskipun demikian, *Half Preprocessing* masih menunjukkan keterbatasannya dalam menghasilkan prediksi yang lebih baik, yang

mengindikasikan perlunya teknik *preprocessing* yang lebih kompleks dan mendalam.

#### 4.5. Analisis

Berdasarkan hasil uji coba yang dilakukan pada berbagai skenario dengan pembagian data 80/20 dan 70/30 serta dua jenis *preprocessing* yaitu *Full Preprocessing* dan *Half Preprocessing*, dapat dianalisis bahwa penggabungan beberapa teknik *preprocessing* memberikan hasil yang lebih optimal dalam hal Akurasi, *F1 Score*, Presisi, dan *Recall*. Tabel 4.5 menyajikan hasil keseluruhan dari uji skenario yang dilakukan.

Tabel 4. 5 Hasil Percobaan Uji Skenario

Skenario Uji	Pembagian Data	Akurasi(%)	<i>F1 Score</i> (%)	Presisi (%)	<i>Recall</i> (%)
<i>Baseline - Full</i>	80/20	42.86	42.99	43.77	42.86
<i>Baseline - Full</i>	70/30	30.43	27.87	28.12	30.00
<i>Baseline - Half</i>	80/20	35.06	35.19	41.83	35.06
<i>Baseline - Half</i>	70/30	24.34	15.44	11.32	24.34
<i>Data Augmentation - Full</i>	80/20	58.82	58.67	61.38	58.82
<i>Data Augmentation - Full</i>	70/30	52.40	52.41	54.51	52.40
<i>Data Augmentation - Half</i>	80/20	44.44	35.06	48.32	44.44

<i>Data Augmentation - Half</i>	70/30	34.06	29.09	47.03	34.06
ROS - Full	80/20	50.51	50.17	55.92	50.51
ROS - Full	70/30	40.94	37.95	38.50	38.00
ROS - Half	80/20	46.46	44.74	44.35	46.46
ROS - Half	70/30	36.91	34.98	40.13	36.50
<i>Data Augmentation + ROS - Full</i>	80/20	73.73	73.37	75.51	73.73
<i>Data Augmentation + ROS - Full</i>	70/30	71.38	71.35	72.24	71.38
<i>Data Augmentation + ROS - Half</i>	80/20	64.14	63.22	63.29	64.14
<i>Data Augmentation + ROS - Half</i>	70/30	56.22	56.11	57.56	56.22

#### 4.5.1. Analisis untuk *Full Preprocessing*

Pada percobaan dengan *Full Preprocessing*, hasil yang diperoleh menunjukkan performa yang lebih baik dibandingkan dengan *Half Preprocessing* di seluruh skenario uji. Pada pembagian data 80/20, model dengan *Full Preprocessing* mencatat Akurasi 73.73% dan *F1 Score* 73.37%, yang menunjukkan kinerja yang sangat baik. Penerapan teknik *Data Augmentation + ROS* pada *Full Preprocessing* memberikan kontribusi signifikan dalam meningkatkan distribusi data yang lebih merata dan memperbaiki kinerja model secara keseluruhan. Hal ini sejalan dengan hasil yang tercatat pada Gambar 4.18, yang menggambarkan distribusi prediksi yang lebih baik pada kebanyakan dimensi kepribadian.

Pada pembagian data 70/30, meskipun ada penurunan performa yang cukup terlihat, model dengan *Full Preprocessing* tetap memberikan hasil yang cukup optimal, dengan Akurasi 71.38% dan *F1 Score* 71.35%. Meskipun ukuran data pelatihan yang lebih kecil mempengaruhi kemampuan model dalam mengenali pola, hasil ini tetap menunjukkan kestabilan model yang lebih baik jika dibandingkan dengan *Half Preprocessing*. Hasil ini dapat dilihat pada Gambar 4.19, yang menunjukkan distribusi prediksi yang lebih merata meskipun dengan data yang terbatas.

#### **4.5.2. Analisis untuk *Half Preprocessing***

Di sisi lain, penggunaan *Half Preprocessing* menunjukkan performa yang lebih rendah dibandingkan dengan *Full Preprocessing* pada semua skenario uji. Meskipun pada pembagian data 80/20 model dengan *Half Preprocessing* menunjukkan peningkatan kinerja dibandingkan dengan *Baseline*, dengan Akurasi 46.46% dan *F1 Score* 44.74%, hasil ini tetap lebih rendah jika dibandingkan dengan *Full Preprocessing* yang mencapai Akurasi 50% lebih tinggi. Hal ini menunjukkan bahwa meskipun *Half Preprocessing* dapat memberikan kontribusi positif, keterbatasan dalam *preprocessing data* yang lebih mendalam menghalangi kemampuan model untuk mengenali pola dengan lebih baik.

Pada pembagian data 70/30, performa model dengan *Half Preprocessing* mengalami penurunan yang cukup signifikan. Dengan Akurasi 36.91% dan *F1 Score* 34.98%, hasil ini menunjukkan bahwa dengan ukuran data pelatihan yang lebih kecil dan teknik *preprocessing* yang tidak terlalu komprehensif, model kesulitan dalam mengenali pola yang ada. Gambar 4.17 menggambarkan distribusi prediksi yang lebih tidak merata, dengan kesalahan yang lebih tinggi pada dimensi *Agreeableness*, *Neuroticism*, dan *Extraversion*, yang menyoroti keterbatasan dari *preprocessing* ini dalam meningkatkan akurasi prediksi.

#### **4.5.3. Perbandingan Kinerja antara *Data Split***

Berdasarkan hasil uji yang dilakukan, perbandingan antara pembagian data 80/20 dan 70/30 menunjukkan bahwa pembagian 80/20 memberikan hasil yang lebih stabil dan lebih baik dalam hal Akurasi, *F1 Score*, Presisi, dan *Recall*. Sebagai contoh, pada *Data Augmentation* + ROS dengan *Full Preprocessing*, pembagian

data 80/20 mencatatkan Akurasi 73.73%, sedangkan pembagian data 70/30 meskipun tetap baik, menunjukkan penurunan menjadi 71.38%. Penurunan ini dapat dijelaskan oleh jumlah data pelatihan yang lebih kecil pada skenario 70/30, yang membatasi kemampuan model untuk mengenali pola secara optimal.

Perbandingan ini juga berlaku pada teknik ROS dan *Data Augmentation* dengan *Full Preprocessing*, di mana pembagian data 80/20 selalu memberikan hasil yang lebih baik dibandingkan dengan pembagian data 70/30. Ini menunjukkan bahwa pembagian data yang lebih besar memberikan peluang lebih bagi model untuk melakukan pelatihan dengan lebih optimal.

Meskipun *Half Preprocessing* menunjukkan penurunan performa pada 70/30, teknik ini lebih stabil pada 80/20, meskipun hasilnya tetap lebih rendah dibandingkan dengan *Full Preprocessing*. Oleh karena itu, dapat disimpulkan bahwa pembagian data 80/20 lebih optimal untuk menghasilkan model yang dapat mengenali pola secara lebih baik, sementara 70/30 berisiko mengurangi akurasi dan ketepatan prediksi pada sebagian besar skenario percobaan.

## BAB 5 KESIMPULAN DAN SARAN

### 5.1. Kesimpulan

Penelitian ini berhasil mengimplementasikan metode RoBERTa untuk klasifikasi kepribadian Big Five berbasis data media sosial X. Hasil terbaik diperoleh dengan kombinasi *Data Augmentation* dan *Random Oversampling* (ROS), yang menghasilkan akurasi 73,73%, presisi 75,51%, recall 73,73%, dan F1-score 73,37% pada skenario pembagian data 80/20 dengan *Full Preprocessing*. Teknik *Full Preprocessing*, yang melibatkan langkah-langkah seperti *normalization*, penghapusan *stopword* dan *stemming*, terbukti memberikan hasil yang lebih baik dibandingkan dengan *Half Preprocessing*, yang lebih sederhana. Namun, keterbatasan jumlah *dataset*, dengan hanya 381 akun, menjadi tantangan signifikan, terutama dalam memprediksi dimensi kepribadian yang lebih sulit dikenali.

### 5.2. Saran

Penelitian selanjutnya disarankan untuk menggunakan *dataset* yang lebih besar agar hasil yang diperoleh lebih representatif terutama dalam meningkatkan akurasi prediksi pada dimensi kepribadian yang lebih jarang atau kurang terwakili. Teknik *preprocessing* juga dapat diperluas dengan metode yang lebih kompleks untuk menangani berbagai karakteristik data yang ada. Selain itu, eksplorasi metode *oversampling* lain seperti SMOTE atau pendekatan berbasis generatif dapat membantu mengatasi ketidakseimbangan data. Terakhir, penggunaan teknik *hyperparameter tuning* yang lebih terkini, seperti *Bayesian Optimization*, dapat dieksplorasi untuk mengoptimalkan proses pelatihan dan mencapai hasil yang lebih maksimal.

## DAFTAR PUSTAKA

- [1] Annur, C.M. (2023) Ada 27 Juta Pengguna Twitter di Indonesia, Terbanyak ke-4 Global. Available at: <https://databoks.katadata.co.id/datapublish/2023/11/28/ada-27-juta-pengguna-twitter-di-indonesia-terbanyak-ke-4-global> (Accessed: 27 March 2024).
- [2] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad, and M. S. Siddik, "Personality Detection from Text using Convolutional Neural Network," in *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*, Institute of Electrical and Electronics Engineers Inc., May 2019. doi: 10.1109/ICASERT.2019.8934548.
- [3] G. D. Salsabila and E. B. Setiawan, "Semantic Approach for Big Five Personality Prediction on Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 680–687, Aug. 2021, doi: 10.29207/resti.v5i4.3197.
- [4] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, [Online]. Available: <http://arxiv.org/abs/1907.11692>.
- [5] M. M. Zain, R. N. Simbolon, H. Sulung, and D. Z. Anwar, "Jurnal Politeknik Caltex Riau Analisis Sentimen Pendapat Masyarakat Mengenai Vaksin Covid-19 Pada Media Sosial Twitter dengan Robustly Optimized BERT Pretraining Approach," 2021. [Online]. Available: <https://jurnal.pcr.ac.id/index.php/jkt/>.
- [6] E. F. Tsani and D. Suhartono, "Personality Identification from Social Media Using Ensemble BERT and RoBERTa," *Informatica (Slovenia)*, vol. 47, no. 4, pp. 537–544, Dec. 2023, doi: 10.31449/inf.v47i4.4771.
- [7] Muh. A. A. Yani and W. Maharani, "Analyzing Cyberbullying Negative Content on Twitter Social Media with the RoBERTa Method," *JINAV: Journal of Information and Visualization*, vol. 4, no. 1, pp. 61–69, Feb. 2023, doi: 10.35877/454ri.jinav1543.
- [8] D. Tiyasya Putra and E. Budi Setiawan, "Sentiment Analysis on Social Media with Glove Using Combination CNN and RoBERTa," vol. 10, no. 3, pp. 457–563, 2023, doi: 10.29207/resti.v7iX3.4892.

- [9] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Detection and Classification of mental illnesses on social media using RoBERTa," Nov. 2020, [Online]. Available: <http://arxiv.org/abs/2011.11226>.
- [10] V. R, "Comparative Analysis for Personality Prediction by Digital Footprints in Social Media," *Journal of Information Technology and Digital World*, vol. 3, no. 2, pp. 77–91, May 2021, doi: 10.36548/jitdw.2021.2.002.
- [11] C. Yuan, J. Wu, H. Li, and L. Wang, "Personality Recognition Based on User Generated Content," in *2018 15th International Conference on Service Systems and Service Management (ICSSSM)*, 2018, pp. 1–6. doi: 10.1109/ICSSSM.2018.8465006.
- [12] J. Xue, A. Li, X. Liu, and L. Cai, "Identifying Big Five personality traits based on facial behavior analysis."
- [13] A. Yilmaz, A. A. Demircali, S. Kocaman, and H. Uvet, "Comparison of Deep Learning and Traditional Machine Learning Techniques for Classification of Pap Smear Images," Sep. 2020, [Online]. Available: <http://arxiv.org/abs/2009.06366>.
- [14] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, doi: <https://doi.org/10.18653/v1/d19-1670>.
- [15] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data: Review of Methods and Applications," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012077, Mar. 2021, doi: <https://doi.org/10.1088/1757-899x/1099/1/012077>
- [16] N. Febrianto, I. Prasetya, and A. Wijaya, "Pembuatan Sistem Prediksi Kepribadian 'The Big Five Traits' dari Media Sosial Twitter." [Online]. Available: [http://semicast.com/en/publications/2012\\_07\\_30\\_Twitter\\_reaches\\_half\\_a\\_billion](http://semicast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion)  
 —.

- [17] M. R. Maarif, “Content Analysis on Twitter Users Interaction within First 100 Days of Jakarta’s New Government by Using Text Mining,” *Journal Pekommas*, vol. 3, no. 2, p. 137, Oct. 2018, doi: 10.30818/jpkm.2018.2030203.
- [18] M. A. Djamaludin, A. Triayudi, and E. Mardiani, “Analisis Sentimen Tweet KRI Nanggala 402 di Twitter menggunakan Metode Naïve Bayes Classifier,” *Jurnal Teknologi Informasi dan Komunikasi*, vol. 6, no. 2, p. 2022, 2022, doi: 10.35870/jti.
- [19] F. Nur, Hasan Dwi Cahyono, Sari Widya Sihwi, and Wisnu Widiarto, “Performance Analysis of Transformer Based Models (BERT, ALBERT, and RoBERTa) in Fake News Detection,” Nov. 2023, doi: <https://doi.org/10.1109/icoiact59844.2023.10455849>.
- [20] M. Stankevich, I. V. Smirnov, Nikolay Ignatiev, O. Grigoryev, and N. Kiselnikova, “Analysis of Big Five Personality Traits by Processing of Social Media Users Activity Features.,” *DAMDID/RCDL*, pp. 162–166, Jan. 2018.
- [21] W. Husain *et al.*, “Reliability generalization meta-analysis of the internal consistency of the Big Five Inventory (BFI) by comparing BFI (44 items) and BFI-2 (60 items) versions controlling for age, sex, language factors,” *BMC Psychology*, vol. 13, no. 1, Jan. 2025, doi: <https://doi.org/10.1186/s40359-024-02271-x>.
- [22] K. Md. Hasib *et al.*, “A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem,” *Journal of Computer Science*, vol. 16, no. 11, pp. 1546–1557, Nov. 2020, doi: <https://doi.org/10.3844/jcssp.2020.1546.1557>.