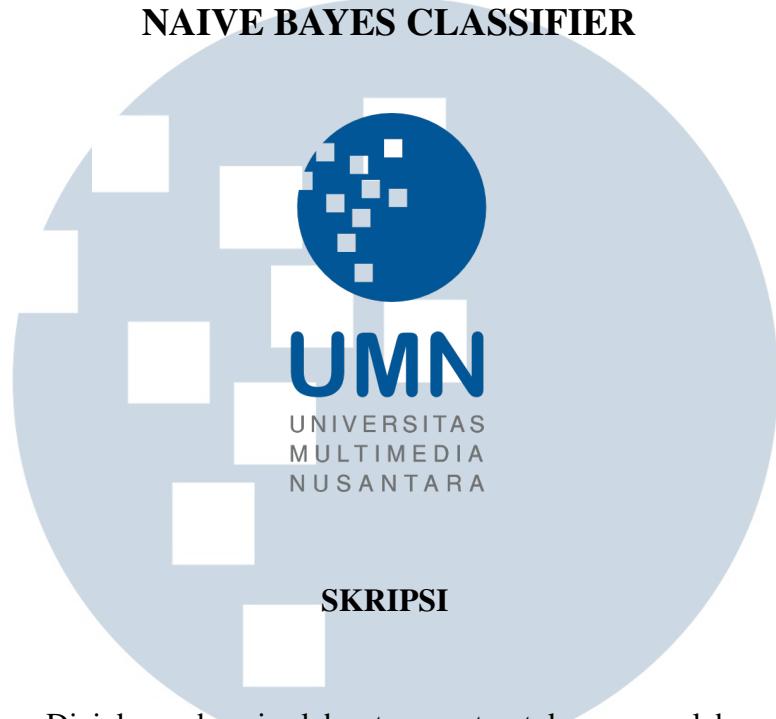


**PENGIDENTIFIKASIAN KALIMAT SARA PADA RUANG
OBROLAN DISCORD MENGGUNAKAN ALGORITMA
NAIVE BAYES CLASSIFIER**



Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer (S.Kom.)

Regina Fransisca Louisa
00000029656

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA
PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2022

**PENGIDENTIFIKASIAN KALIMAT SARA PADA RUANG OBROLAN
DISCORD MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER**



Regina Fransisca Louisa
00000029656

UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG
2022

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Regina Fransisca Louisa
Nomor Induk Mahasiswa : 00000029656
Program Studi : Informatika

Skripsi dengan judul:

Pengidentifikasi Kalimat SARA Pada Ruang Obrolan Discord Menggunakan Algoritma Naive Bayes Classifier

merupakan hasil karya saya sendiri bukan plagiat dari karya ilmiah yang ditulis oleh orang lain, dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/ penyimpangan, baik dalam pelaksanaan Skripsi maupun dalam penulisan laporan Skripsi, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk Tugas akhir yang telah saya tempuh.

Tangerang, 17 Juni 2022



(Regina Fransisca Louisa)

**UNIVERSITAS
MULTIMEDIA
NUSANTARA**

HALAMAN PENGESAHAN

Skripsi dengan judul

PENGIDENTIFIKASIAN KALIMAT SARA PADA RUANG OBROLAN DISCORD MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER

oleh

Nama : Regina Fransisca Louisa
NIM : 00000029656
Program Studi : Informatika
Fakultas : Fakultas Teknik dan Informatika

Telah diujikan pada hari Selasa, Tgl. 28 Juni 2022

Pukul 13.00 s/s 15.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut

Ketua Sidang

Penguji

(Dennis Gunawan, S.Kom., M.Sc.)

NIDN: 0320059001

Pembimbing I

(Farica Perdana Putri, S.Kom., M.Sc.)

NIDN: 0331019301

Pembimbing II

(Moeljono Widjaja, B.Sc.,M.Sc.,Ph.D.)

NIDN: 0311106903

(Marlinda Vasty Overbeek, S.Kom.,
M.Kom.)

NIDN: 0818038501

Ketua Program Studi Informatika,

(Marlinda Vasty Overbeek, S.Kom., M.Kom.)

NIDN: 0818038501

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Multimedia Nusantara, saya yang bertanda tangan di bawah ini:

Nama	:	Regina Fransisca Louisa
NIM	:	00000029656
Program Studi	:	Informatika
Fakultas	:	Teknik dan Informatika
Jenis Karya	:	Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada **Universitas Multimedia Nusantara** hak Bebas Royalti Non-eksklusif (*Non-exclusive Royalty-Free Right*) atas karya ilmiah saya yang berjudul:

PENGIDENTIFIKASIAN KALIMAT SARA PADA RUANG OBROLAN DISCORD MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non eksklusif ini Universitas Multimedia Nusantara berhak menyimpan, mengalih media / format-kan, mengelola dalam bentuk pangkalan data (*database*), merawat, dan mempublikasikan tugas akhir saya selama tetap mencantumkan nama saya sebagai penulis / pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenarnya.

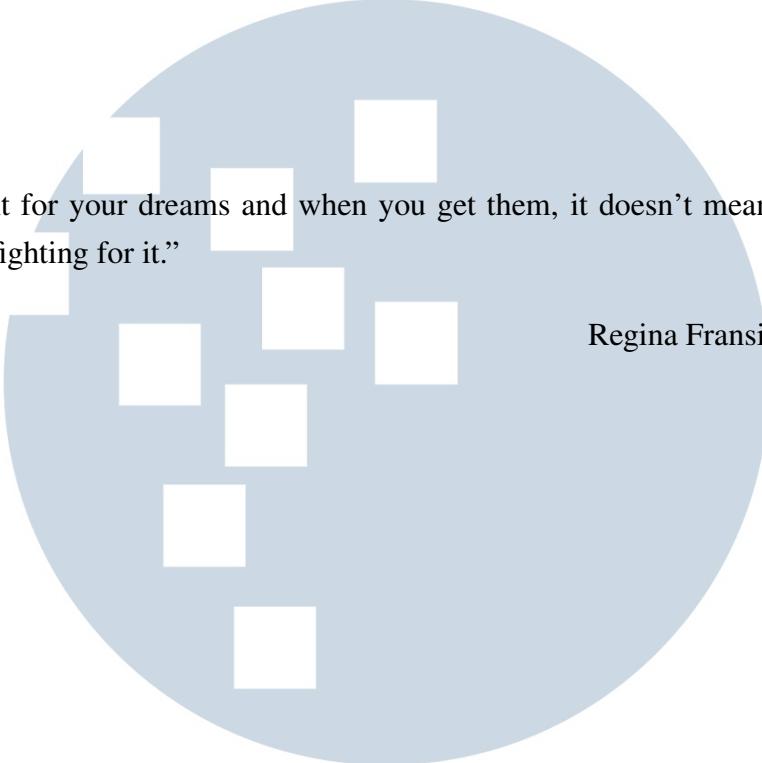
Tangerang, 17 Juni 2022

Yang menyatakan



Regina Fransisca Louisa

Halaman Persembahan / Motto



”Fight for your dreams and when you get them, it doesn’t mean you stop fighting for it.”

Regina Francisca Louisa

UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

KATA PENGANTAR

Puji Syukur atas berkat dan rahmat kepada Tuhan Yang Maha Esa, atas selesainya penulisan laporan Skripsi ini dengan judul: Pengidentifikasi Kalimat SARA Pada Ruang Obrolan Discord Menggunakan Algoritma Naive Bayes Classifier dilakukan untuk memenuhi salah satu syarat untuk mencapai gelar Sarjana Komputer Jurusan Informatika Pada Fakultas Teknik dan Informatika Universitas Multimedia Nusantara. Menyadari bahwa tanpa bantuan dan bimbingan dari berbagai pihak, dari masa perkuliahan sampai pada penyusunan skripsi ini, sangatlah sulit bagi untuk menyelesaikan skripsi ini. Ucapan terima kasih diberikan kepada:

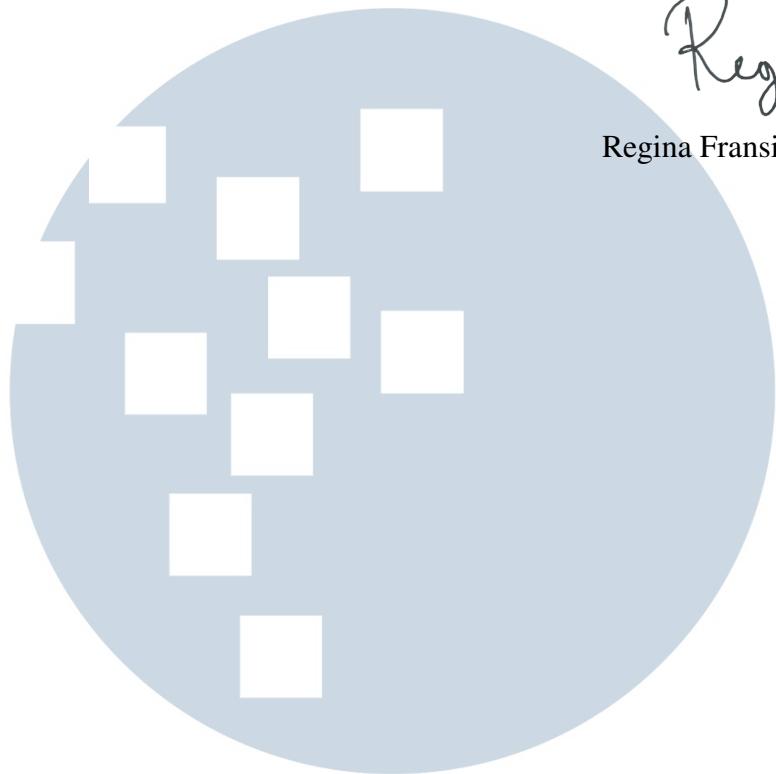
1. Tuhan yang Maha Esa.
2. Orangtua dan keluarga yang telah memberikan bantuan dukungan material dan moral, sehingga Skripsi ini dapat diselesaikan dengan baik.
3. Bapak Dr. Ninok Leksono, selaku Rektor Universitas Multimedia Nusantara.
4. Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
5. Ibu Marlinda Vasty Overbeek, S.Kom., M.Kom., selaku Ketua Program Studi Informatika Universitas Multimedia Nusantara dan sebagai Pembimbing kedua yang telah banyak membantu dan memberikan bimbingan hingga Skripsi ini dapat diselesaikan dengan baik.
6. Bapak Moeljono Widjaja, B.Sc.,M.Sc.,Ph.D., sebagai Pembimbing pertama yang telah banyak meluangkan waktu untuk memberikan bimbingan, arahan dan motivasi hingga Skripsi ini dapat diselesaikan dengan baik.
7. Teman-teman kuliah yang memberikan semangat dan motivasi saat menjalankan kuliah bersama dari awal hingga saat ini.

Semoga skripsi ini bermanfaat, baik sebagai sumber informasi maupun sumber inspirasi, bagi para pembaca.

Tangerang, 17 Juni 2022

Regina

Regina Francisca Louisa



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

PENGIDENTIFIKASIAN KALIMAT SARA PADA RUANG OBROLAN DISCORD MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER

Regina Fransisca Louisa

ABSTRAK

Discord merupakan aplikasi media sosial berbasis iOS, Windows, dan Android yang didesain khusus untuk mendukung para pemain *game* untuk dapat berkomunikasi pada saat bermain game. Berdasarkan hasil survei yang dilakukan melalui Google Form pada 24 hingga 26 April 2022, didapatkan hasil bahwa terdapat banyak permasalahan yang disebabkan oleh kalimat SARA pada Discord. Dibutuhkan penelitian untuk mengidentifikasi apakah kalimat pada ruang obrolan Discord terindikasi sebagai SARA atau bukan. Untuk dapat mengetahui bahwa suatu kalimat dalam obrolan discord merupakan kalimat yang mengandung SARA atau bukan, diperlukan sebuah sistem yang dapat mengklasifikasikan data. Penelitian dilakukan dengan menggunakan algoritma Naive Bayes Classifier. Naive Bayes Classifier merupakan algoritma yang digunakan untuk melakukan klasifikasi suatu data dengan metode probabilitas dan statistik. Hasil penelitian dapat dievaluasi berdasarkan tingkat *accuracy*, *precision*, *recall*, dan *f1-score*. Dalam pengidentifikasi kalimat SARA pada ruang obrolan Discord, dihasilkan tingkat *accuracy* 94.25%, *precision* 0.98, *recall* 0.91, dan *f1-score* 0.95. Berdasarkan hasil penelitian, sistem berhasil melakukan identifikasi kalimat SARA pada Discord.

Kata kunci: *Accuracy*, *Discord*, *F1-Score*, *Naive Bayes Classifier*, *Precision*, *Recall*



Identification of SARA Sentences in Discord Chat Room Using Naive Bayes Classifier Algorithm

Regina Fransisca Louisa

ABSTRACT

Discord is a social media application based on iOS, Windows, and Android which is specially designed to support gamers to communicate while playing games. Based on the results of a survey conducted through Google Form from April 24 to 26, 2022, there were so many problems caused by sentences containing elements of ethnicity, religion, racism, and intergroup on Discord. The research was made to identify sentences in Discord chat room indicated as sentences containing ethnicity, religion, racism, and inter-group or not. The research was conducted using the Naive Bayes Classifier algorithm. Naive Bayes Classifier is an algorithm used to classify data with probability and statistical methods. The results of the research can be evaluated based on the levels of accuracy, precision, recall, and f1-score. In identifying sentences containing elements of ethnicity, religion, racism, and inter-group in the Discord chat room, the result of the level of accuracy was 94.25%, the result of the level of precision was 0.98, the result of the level of recall was 0.91, and the result of the level of f1-score was 0.95. Based on the results, the system successfully identified sentences that contained ethnic, religion, racism, and inter-group on Discord.

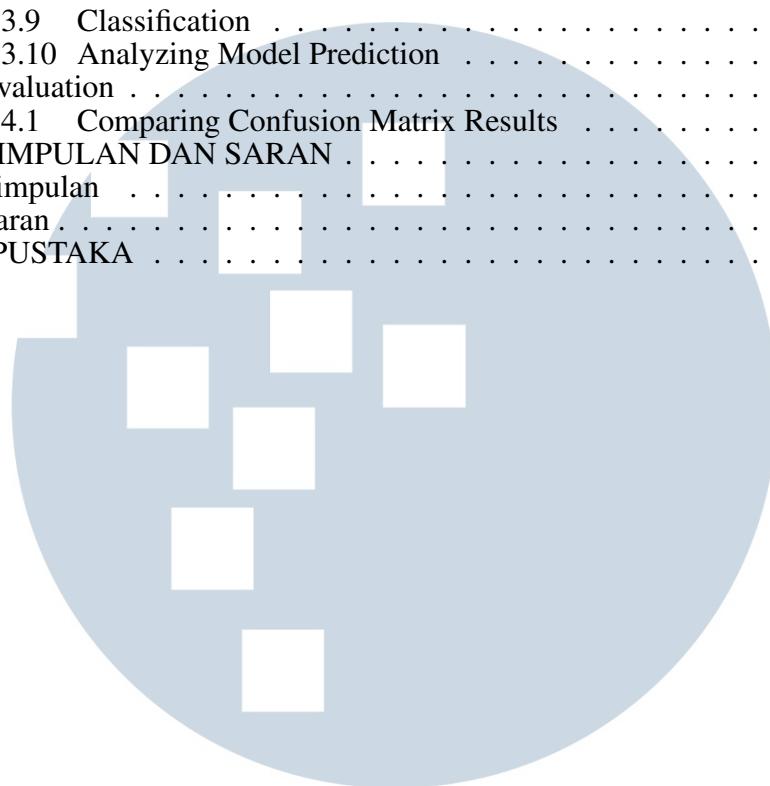
Keywords: Accuracy, Discord, F1-Score, Naive Bayes Classifier, Precision, Recall



DAFTAR ISI

HALAMAN JUDUL	i
PERNYATAAN TIDAK MELAKUKAN PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI ILMIAH	iv
HALAMAN PERSEMBAHAN/MOTO	v
KATA PENGANTAR	vi
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI	x
DAFTAR GAMBAR	xii
DAFTAR TABEL	xiii
DAFTAR KODE	xiv
DAFTAR LAMPIRAN	xv
BAB 1 PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Rumusan Masalah	3
1.3 Batasan Permasalahan	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian	4
1.6 Sistematika Penulisan	4
BAB 2 LANDASAN TEORI	6
2.1 Pembelajaran Mesin	6
2.2 Naïve Bayes Classifier	7
2.3 SARA	8
2.4 Confusion Matrix	8
2.5 Discord	10
BAB 3 METODOLOGI PENELITIAN	12
3.1 Metodologi Penelitian	12
3.1.1 Scraping Data	13
3.1.2 Labeling	14
3.1.3 Tahap Preprocessing	14
3.1.4 Train-test Split	17
3.1.5 Feature Extraction	17
3.1.6 Classification	17
3.1.7 Analyzing Model Prediction	17
3.1.8 Evaluation	18
BAB 4 HASIL DAN DISKUSI	19
4.1 System Specifications	19
4.2 Data Collection	19
4.2.1 Scraping Data	20
4.2.2 Labeling	22
4.3 Build Classification System	23
4.3.1 Load Data	23
4.3.2 Remove Unused Column	24
4.3.3 Check The Amount of Data	24
4.3.4 Convert The Label Values	25
4.3.5 Tahap Preprocessing	26
4.3.6 Train-test Split	33

4.3.7	Handle The Imbalance Data	34
4.3.8	Feature Extraction	40
4.3.9	Classification	40
4.3.10	Analyzing Model Prediction	40
4.4	Evaluation	41
4.4.1	Comparing Confusion Matrix Results	42
BAB 5	SIMPULAN DAN SARAN	54
5.1	Simpulan	54
5.2	Saran	54
DAFTAR PUSTAKA	DAFTAR PUSTAKA	55



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR GAMBAR

Gambar 3.1	Tahapan penelitian	13
Gambar 3.2	Proses <i>preprocessing</i>	14
Gambar 3.3	Proses <i>case folding</i>	15
Gambar 3.4	Proses <i>cleaning</i>	15
Gambar 3.5	Proses <i>tokenizing</i>	15
Gambar 3.6	Proses <i>slangword removal</i>	16
Gambar 3.7	Proses <i>stopwords removal</i>	16
Gambar 3.8	Proses <i>stemming</i>	16
Gambar 3.9	Proses <i>countvectorizer</i>	17
Gambar 4.1	Tahapan <i>scraping data</i>	20
Gambar 4.2	Tahapan <i>copy-paste the text-channel's link</i>	21
Gambar 4.3	Tahapan <i>copy-paste Discord invitation link</i>	21
Gambar 4.4	Tahapan <i>copy-paste Discord token id</i>	22
Gambar 4.5	<i>Scraping data success</i>	22
Gambar 4.6	<i>Labeling</i>	23
Gambar 4.7	<i>Imbalance Data</i>	25
Gambar 4.8	<i>Case folding result</i>	26
Gambar 4.9	<i>Data cleaning result</i>	28
Gambar 4.10	<i>Tokenizing result</i>	29
Gambar 4.11	<i>Remove slangwords result</i>	30
Gambar 4.12	<i>Remove stopwords result</i>	32
Gambar 4.13	<i>Stemming result</i>	33
Gambar 4.14	Hasil bar plot dari <i>training data</i> dengan teknik <i>oversampling</i>	37
Gambar 4.15	Hasil bar plot dari <i>testing data</i> dengan teknik <i>oversampling</i>	39
Gambar 4.16	<i>Train-test label bar plot</i> berdasarkan hasil teknik <i>oversampling</i>	39
Gambar 4.17	Hasil prediksi dengan sembarang data	41
Gambar 4.18	<i>Train-test label bar plot</i> tanpa penyeimbangan data	42
Gambar 4.19	<i>Confusion matrix</i> tanpa dilakukan penyeimbangan data	44
Gambar 4.20	<i>Confusion matrix</i> tanpa <i>preprocessing</i>	46
Gambar 4.21	Hasil bar plot dari <i>training dataset</i> dengan teknik <i>undersampling</i>	48
Gambar 4.22	Hasil bar plot dari <i>testing data</i> dengan teknik <i>undersampling</i>	49
Gambar 4.23	<i>Train-test label bar plot</i> berdasarkan hasil teknik <i>undersampling</i>	50
Gambar 4.24	<i>Confusion matrix</i> dengan teknik <i>undersampling</i>	51
Gambar 4.25	<i>Confusion matrix</i>	53

DAFTAR TABEL

Tabel 2.1	Tabel Confusion Matrix	9
Tabel 4.1	Tabel dataset	23
Tabel 4.2	Tabel hasil <i>case folding</i> , <i>data cleaning</i> , dan <i>tokenizing</i> . . .	29
Tabel 4.3	Tabel <i>classification report</i> tanpa dilakukan penyeimbangan data	43
Tabel 4.4	Tabel <i>classification report</i> tanpa <i>preprocessing</i>	45
Tabel 4.5	Tabel <i>classification report</i> dengan teknik <i>undersampling</i> . .	50
Tabel 4.6	Tabel <i>classification report</i> dengan teknik <i>oversampling</i> . .	52



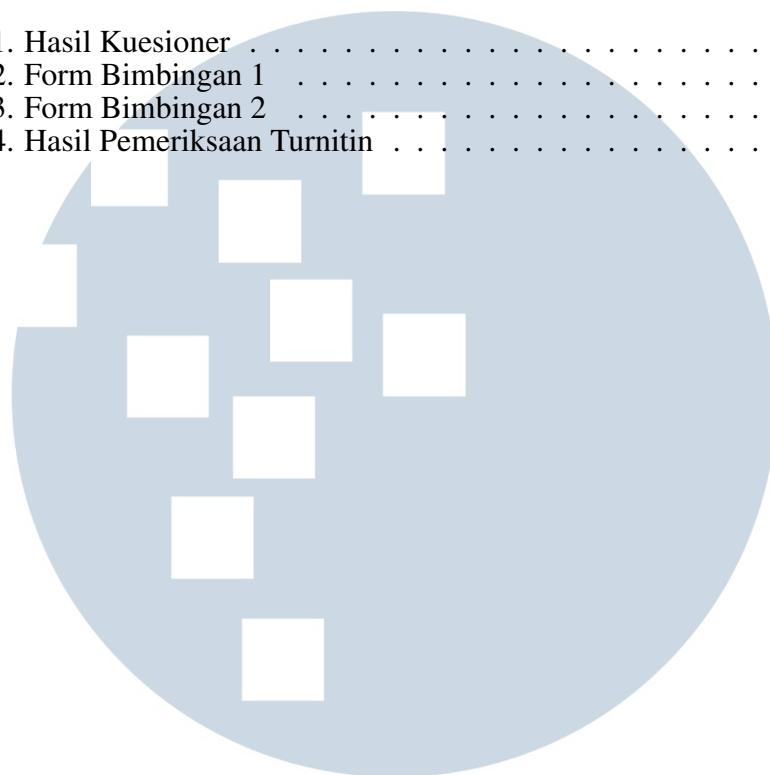
DAFTAR KODE

4.1	Potongan kode untuk menghilangkan kolom yang tidak digunakan	24
4.2	Potongan kode proses perubahan nilai label	25
4.3	Potongan kode proses <i>case folding</i>	26
4.4	Potongan kode proses <i>data cleaning</i>	27
4.5	Potongan kode <i>tokenizing</i>	28
4.6	Potongan kode <i>remove slangwords</i>	29
4.7	Potongan kode <i>remove stopwords</i>	31
4.8	Potongan kode <i>stemming</i>	32
4.9	Potongan kode <i>train-test split</i>	33
4.10	Potongan kode penggabungan data menjadi <i>training dataset</i> dan <i>testing dataset</i>	34
4.11	Potongan kode proses pemisahan data berdasarkan kelas	35
4.12	Potongan kode penyeimbangan data pada <i>training dataset</i> dengan teknik <i>oversampling</i>	36
4.13	Potongan kode penyeimbangan data pada <i>testing dataset</i> dengan teknik <i>oversampling</i>	37
4.14	Potongan kode <i>CountVectorizer</i>	40
4.15	Potongan kode <i>classification</i>	40
4.16	Potongan kode prediksi menggunakan data uji	40
4.17	Potongan kode prediksi menggunakan sembarang data	41
4.18	Potongan kode penyeimbangan data pada <i>training dataset</i> dengan teknik <i>undersampling</i>	47
4.19	Potongan kode penyeimbangan data pada <i>testing dataset</i> dengan teknik <i>undersampling</i>	48



DAFTAR LAMPIRAN

Lampiran 1. Hasil Kuesioner	56
Lampiran 2. Form Bimbingan 1	61
Lampiran 3. Form Bimbingan 2	63
Lampiran 4. Hasil Pemeriksaan Turnitin	65



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA