

**ANALISIS SENTIMEN TWITTER TERHADAP PILPRES
INDONESIA 2024 MENGGUNAKAN ALGORITMA NAIVE
BAYES DENGAN TEKNIK UNDERSAMPLING UNTUK
MENGATASI KETIDAKSEIMBANGAN KELAS**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Teknik Komputer



disusun oleh

AKRAM KEMAL DEWANTARA

20.83.0569

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2025

**ANALISIS SENTIMEN TWITTER TERHADAP PILPRES
INDONESIA 2024 MENGGUNAKAN ALGORITMA NAIVE
BAYES DENGAN TEKNIK UNDERSAMPLING UNTUK
MENGATASI KETIDAKSEIMBANGAN KELAS**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Teknik Komputer



disusun oleh

AKRAM KEMAL DEWANTARA

20.83.0569

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2025

HALAMAN PERSETUJUAN

SKRIPSI

**ANALISIS SENTIMEN TWITTER TERHADAP PILPRES
INDONESIA 2024 MENGGUNAKAN ALGORITMA NAIVE
BAYES DENGAN TEKNIK UNDERSAMPLING UNTUK
MENGATASI KETIDAKSEIMBANGAN KELAS**

yang disusun dan diajukan oleh

AKRAM KEMAL DEWANTARA

20.83.0569

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 31 Januari 2025

Dosen Pembimbing,



Jeki Kuswanto, M.Kom.

NIK. 190302456

HALAMAN PENGESAHAN

SKRIPSI

**ANALISIS SENTIMEN TWITTER TERHADAP PILPRES INDONESIA 2024
MENGUNAKAN ALGORITMA NAIVE BAYES DENGAN TEKNIK
UNDERSAMPLING UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS**

yang disusun dan diajukan oleh

AKRAM KEMAL DEWANTARA

20.83.0569

Telah dipertahankan di depan Dewan Penguji
pada tanggal 31 Januari 2025

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Wahid Miftahul Ashari, S.Kom., M.T.
NIK. 190302452

Banu Santoso, S.T., M.Eng.
NIK. 190302327

Jeki Kuswanto, M.Kom.
NIK. 190302456

Skrripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 31 Januari 2025

DEKAN FAKULTAS ILMU KOMPUTER



Hanif Al Fatta, S.Kom., M.Kom., Ph.D.
NIK. 190302096

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : AKRAM KEMAL DEWANTARA
NIM : 20.83.0569

Menyatakan bahwa Skripsi dengan judul berikut:

**ANALISIS SENTIMEN TWITTER TERHADAP PILPRES INDONESIA
2024 MENGGUNAKAN ALGORITMA NAIVE BAYES DENGAN TEKNIK
UNDERSAMPLING UNTUK MENGATASI KETIDAKSEIMBANGAN
KELAS**

Dosen Pembimbing : Jeki Kuswanto, M.Kom.

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

Yogyakarta, 31 Januari 2025

Yang Menyatakan,



Akram Kemal Dewantara

HALAMAN PERSEMBAHAN

Saya mengucapkan rasa terima kasih dan penghormatan yang setinggi-tingginya kepada kedua orang tua saya atas dukungan moral maupun finansial yang tak henti-hentinya diberikan selama menempuh pendidikan ini. Tanpa doa, kasih sayang, dan pengorbanan mereka, pencapaian ini tidak akan mungkin terwujud.

Rasa terima kasih juga saya tujukan kepada seluruh dosen dan staf di Universitas Amikom Yogyakarta, yang telah memberikan ilmu, pengalaman, serta fasilitas yang mendukung studi saya. Terima kasih atas ilmu yang berharga serta kesempatan untuk mengembangkan diri di lingkungan akademik yang penuh dukungan.

Ucapan terima kasih ini juga saya sampaikan kepada teman-teman seperjuangan, yang selalu siap membantu, berbagi suka dan duka, serta saling mendukung di setiap tahap perkuliahan hingga proses penyusunan tugas akhir ini. Kehadiran kalian semua membuat perjalanan ini menjadi lebih bermakna dan berkesan.

Terakhir, penghargaan dan terima kasih kepada semua pihak yang tidak dapat saya sebutkan satu per satu, yang telah memberikan bantuan dan dukungannya, baik secara langsung maupun tidak langsung. Semoga kebaikan dan dukungan kalian mendapat balasan yang berlipat ganda.

Dengan rasa syukur dan rendah hati, karya ini saya persembahkan.

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan penyusunan skripsi dengan judul “ANALISIS SENTIMEN TWITTER TERHADAP PILPRES INDONESIA 2024 MENGGUNAKAN ALGORITMA NAIVE BAYES DENGAN TEKNIK UNDERSAMPLING UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS” dengan baik. Penulis berharap hasil penelitian ini dapat memberikan kontribusi bagi pengembangan ilmu pengetahuan, khususnya di bidang analisis sentimen dan pemrosesan bahasa alami (Natural Language Processing).

Penelitian ini disusun sebagai salah satu syarat untuk menyelesaikan program studi di Teknik Komputer pada Universitas Amikom Yogyakarta. Dalam proses penyusunan skripsi ini, penulis menyadari bahwa skripsi ini tidak akan dapat terselesaikan tanpa bantuan, dukungan, dan bimbingan dari berbagai pihak. Oleh karena itu, dengan penuh rasa hormat, penulis mengucapkan terima kasih kepada:

1. Bapak, Ibu serta Adik saya atas dukungan moral dan kasih sayang yang terus berbalas tiada hentinya
2. Bapak Jeki Kuswanto, M.Kom. yang dengan sabar memberikan arahan, bimbingan, dan motivasi selama proses penyusunan penelitian ini.
3. Rekan TK02 dan Teman-teman seperjuangan saya yang telah memberikan semangat, bantuan, dan kebersamaan yang berarti selama proses penyusunan skripsi ini.
4. *Last but not least, I wanna thank for me, I wanna thank me for believing in me, I wanna thank me for doing all this hard works, “Logic will get you from A to Z; Imagination will get you everywhere.”- Albert Einstein*

Akhir kata, semoga skripsi ini dapat bermanfaat bagi semua pihak yang berkepentingan, serta menjadi langkah kecil yang berarti dalam mengembangkan ilmu pengetahuan.

Yogyakarta, 31 Januari 2025

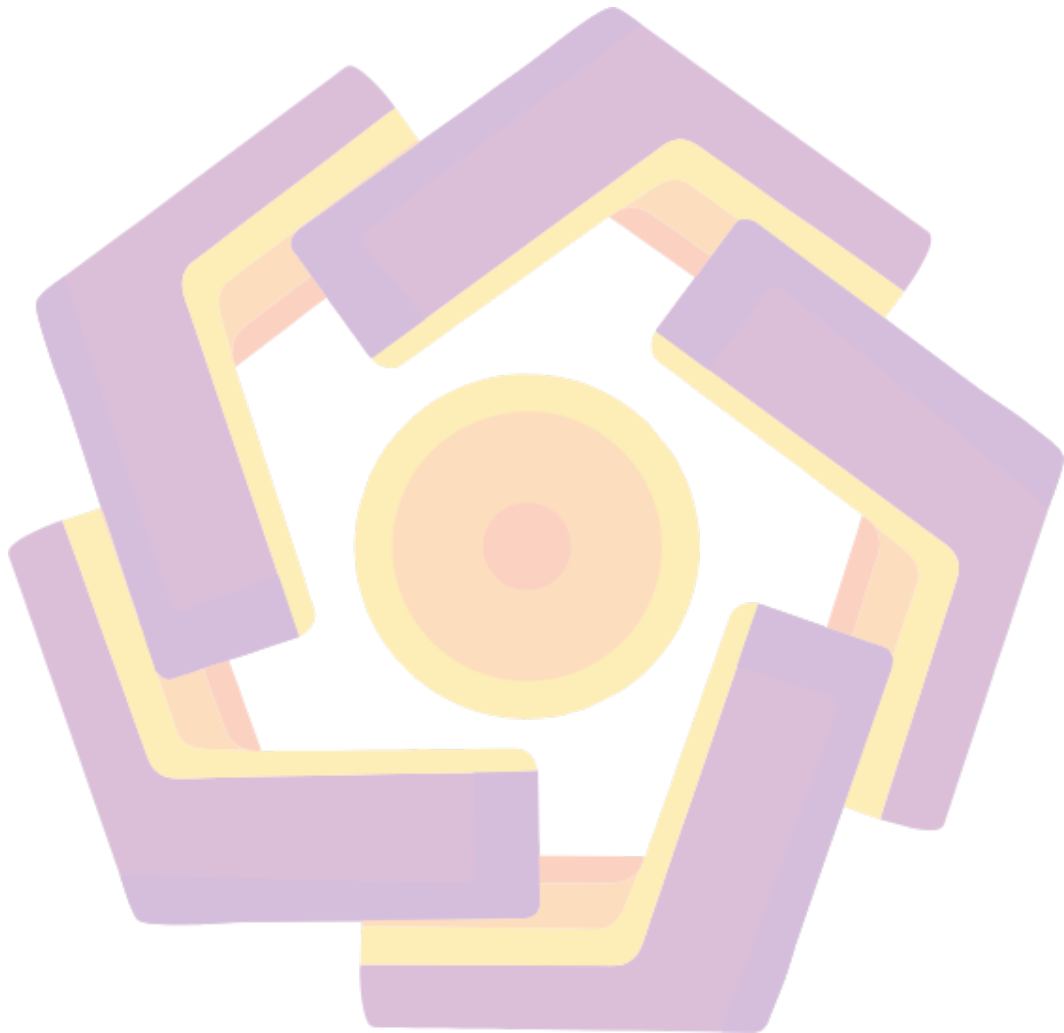
DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN.....	xiii
DAFTAR LAMBANG DAN SINGKATAN	xiv
DAFTAR ISTILAH.....	xv
INTISARI	xvii
ABSTRACT.....	xviii
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah.....	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	3
1.5 Manfaat Penelitian	4
1.6 Sistematika Penulisan	5
BAB II TINJAUAN PUSTAKA	7
2.1 Studi Literatur	7

2.2	Dasar Teori.....	12
2.2.1	Twitter.....	12
2.2.2	Analisis Sentimen	12
2.2.3	Algoritma Naive Bayes.....	13
2.2.4	Teknik Undersampling.....	14
2.2.5	Preprocessing Data.....	15
2.2.6	Splitting Dataset.....	16
2.2.7	Evaluasi Model	17
2.2.8	Term Frequency-Inverse Document Frequency (TF-IDF)	17
2.2.9	Python dan Pustaka Pendukung.....	18
2.2.10	Google Colab	18
BAB III METODE PENELITIAN		20
3.1	Objek Penelitian.....	20
3.2	Alur Penelitian	20
3.3	Data dan Pustaka.....	21
3.3.1	Dataset.....	21
3.3.2	Pustaka dan Perangkat Lunak	22
3.4	Pre-processing Data	22
3.4.1	Case Folding dan Penghapusan Karakter Khusus.....	22
3.4.2	Tokenization	23
3.4.3	Stopword Removal dan Lemmatizing.....	23
3.4.4	Koreksi Penulisan	24
3.4.5	Stemming	24
3.4.6	Hapus teks “NaN”	25
3.5	Labelling dan Undersampling.....	25

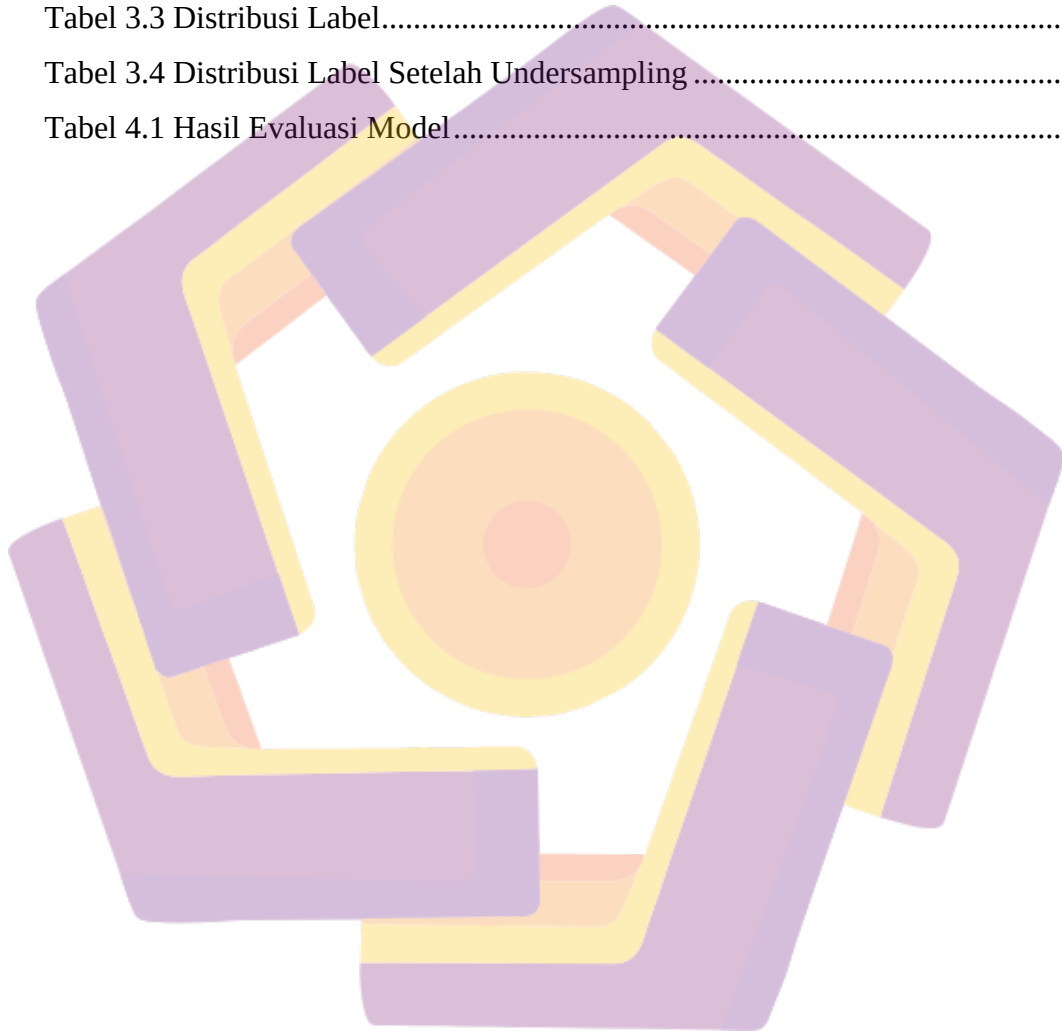
3.5.1	Proses Labeling	26
3.5.2	Undersampling	26
3.5.3	Visualisasi Wordcloud	27
3.6	Splitting Data	28
3.7	Modelling	30
3.8	Evaluasi dan Validasi Model	30
3.8.1	Teknik Evaluasi Model	30
3.8.2	Teknik Validasi Model	32
3.8.3	Perbandingan Akurasi Baseline dan Model	32
BAB IV HASIL DAN PEMBAHASAN		33
4.1	Hasil Pre-Processing Data	33
4.1.1	Distribusi Panjang Teks	33
4.1.2	Pengecekan Nilai Kosong	34
4.2	Hasil Labelling dan Undersampling	34
4.2.1	Hasil Labelling	35
4.2.2	Hasil Undersampling	35
4.2.3	Visualisasi Wordcloud	36
4.3	Hasil Splitting Data	38
4.4	Hasil Modelling	40
4.5	Hasil Evaluasi dan Validasi Model	41
4.5.1	Classification Report	41
4.5.2	Confusion Matrix	42
4.5.3	Hasil K-Fold Cross-Validation	43
4.5.4	Perbandingan Akurasi Baseline dan Model	43
BAB V PENUTUP		45

5.1	Kesimpulan	45
5.2	Saran	45
REFERENSI		46
LAMPIRAN.....		49



DAFTAR TABEL

Tabel 2.1 Keaslian Penelitian	27
Tabel 3.1 Aspek dan Deskripsi Data penelitian	39
Tabel 3.2 Pustaka dan Perangkat Lunak penelitian	40
Tabel 3.3 Distribusi Label.....	44
Tabel 3.4 Distribusi Label Setelah Undersampling	45
Tabel 4.1 Hasil Evaluasi Model.....	59

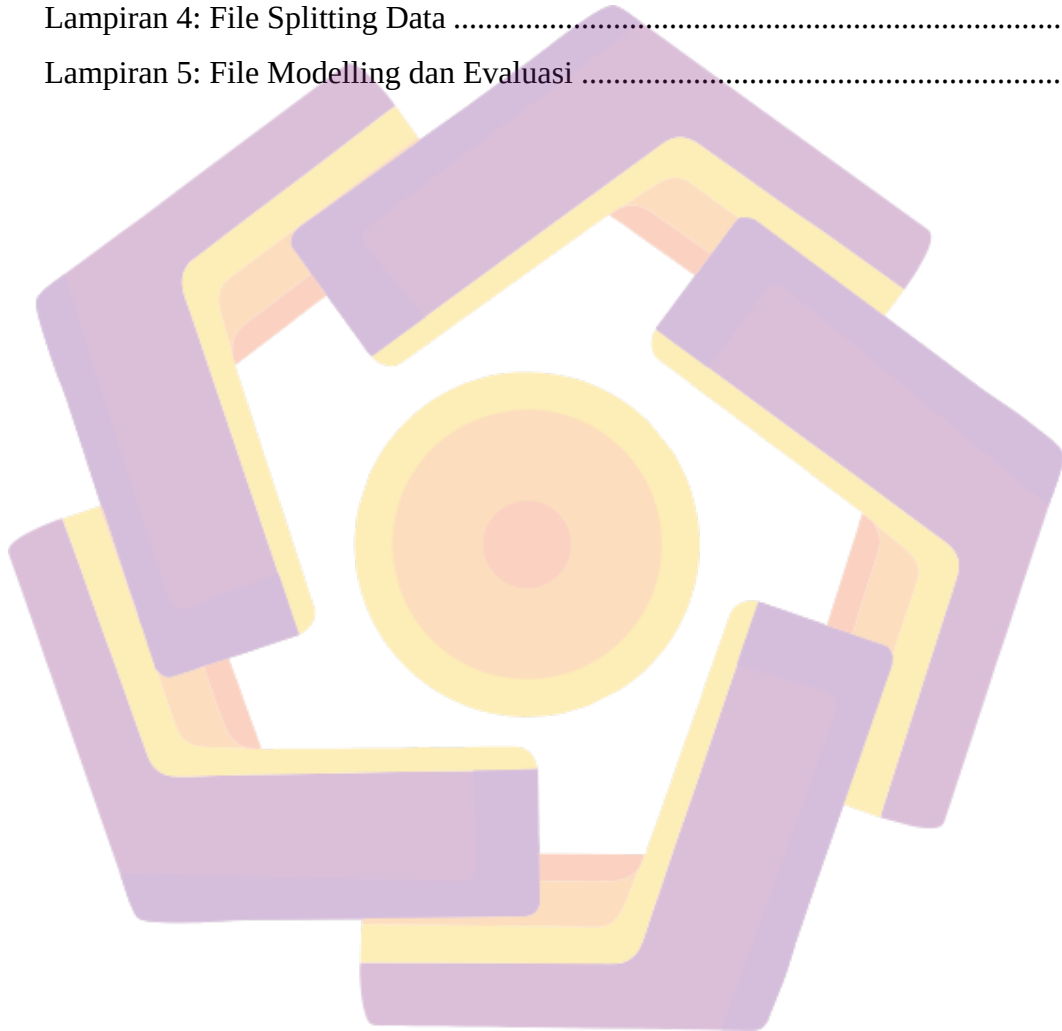


DAFTAR GAMBAR

Gambar 2.1 Twitter	30
Gambar 2.2 Analisis Sentimen	31
Gambar 2.3 Algoritma Naïve Bayes	32
Gambar 2.4 Teknik Undersampling	33
Gambar 2.5 Tahap Preprocessing Data	34
Gambar 2.6 Tahap Splitting Data	34
Gambar 2.7 Tahap Evaluasi Model	35
Gambar 2.8 Teknik TF-IDF	36
Gambar 2.9 Python	36
Gambar 2.10 Google colab	37
Gambar 3.1 Alur Penelitian	39
Gambar 3.2 Case Folding dan Penghapusan Karakter Khusus	41
Gambar 3.3 Tokenization	41
Gambar 3.4 Stopword Removal dan Lemmatizing	42
Gambar 3.5 Koreksi Penulisan	42
Gambar 3.6 Stemming	43
Gambar 4.1 Distribusi Panjang Teks	51
Gambar 4.2 Pengecekan Nilai Kosong	52
Gambar 4.3 Hasil Labelling	53
Gambar 4.4 Hasil Undersampling	54
Gambar 4.5 Keseimbangan Label Sentimen	54
Gambar 4.6 Wordcloud Sentimen Positif	55
Gambar 4.7 Wordcloud Sentimen Negatif	55
Gambar 4.8 Wordcloud Sentimen Netral	56
Gambar 4.9 Visualisasi Splitting Data	57
Gambar 4.10 Hasil Modelling	58
Gambar 4.11 Classification Report	60
Gambar 4.12 Confusion Matrix	60
Gambar 4.13 Hasil K-Fold Cross Validation	61
Gambar 4.14 Perbandingan Akurasi Baseline dan Model	62

DAFTAR LAMPIRAN

Lampiran 1: Import Library dan Dataset	69
Lampiran 2: File Preprocessing Data.....	70
Lampiran 3: File Labelling, Undersampling dan Wordcloud	74
Lampiran 4: File Splitting Data	77
Lampiran 5: File Modelling dan Evaluasi	78



DAFTAR LAMBANG DAN SINGKATAN

μ : Rata-rata (Mean)

σ : Simpangan baku (Standard Deviation)

P : Probabilitas

Ω : Ruang sampel

API : Application Programming Interface

AI : Artificial Intelligence

CNN : Convolutional Neural Network

CSV : Comma Separated Values

GPU : Graphics Processing Unit

JSON : JavaScript Object Notation

ML : Machine Learning

NB : Naive Bayes

NLP : Natural Language Processing

POS : Part-of-Speech

TF-IDF : Term Frequency-Inverse Document Frequency

DAFTAR ISTILAH

Algoritma:	Langkah-langkah logis yang disusun secara sistematis untuk menyelesaikan suatu masalah atau menjalankan suatu tugas tertentu.
Analisis Sentimen:	Proses mengidentifikasi, mengekstraksi, dan mengklasifikasikan sentimen dari teks, biasanya dikategorikan menjadi positif, negatif, atau netral.
Case Folding:	Proses normalisasi teks dengan mengubah semua huruf menjadi huruf kecil untuk menjaga konsistensi data.
Classification Report:	Laporan evaluasi yang mencakup metrik seperti presisi, recall, dan F1-score untuk mengukur performa model klasifikasi.
Confusion Matrix:	Matriks yang menunjukkan perbandingan antara prediksi model dan data aktual untuk evaluasi akurasi model.
Cross Validation:	Teknik pembagian dataset menjadi beberapa subset untuk melatih dan menguji model secara berulang guna meningkatkan keandalan hasil.
Data Unlabeled:	Dataset yang tidak memiliki label atau kategori tertentu sehingga memerlukan pelabelan manual atau otomatis sebelum digunakan.
K-Fold cross validation:	Metode validasi yang membagi data menjadi k subset untuk memastikan model diuji secara menyeluruh pada berbagai bagian data.
Lemmatization:	Proses mengubah kata-kata ke bentuk dasar dengan mempertimbangkan konteksnya.
Machine Learning:	Bidang kecerdasan buatan yang memungkinkan komputer belajar dari data tanpa diprogram secara eksplisit.

Multinomial Naive Bayes (MultinomialNB):

Salah satu algoritma Naive Bayes yang sering digunakan untuk klasifikasi teks berbasis probabilitas.

- Preprocessing: Serangkaian proses yang bertujuan membersihkan dan mempersiapkan data agar lebih sesuai untuk analisis atau pemodelan.
- Precision: Metrik evaluasi yang mengukur proporsi prediksi benar dari semua prediksi positif.
- Recall: Metrik evaluasi yang mengukur proporsi data sebenarnya positif yang berhasil teridentifikasi oleh model.
- Residu Kesalahan: Data atau nilai yang tidak sesuai dengan prediksi atau hasil analisis model.
- Stemming: Proses mengubah kata menjadi akar katanya dengan menghilangkan imbuhan.
- Stopword Removal: Proses menghapus kata-kata umum yang tidak memberikan kontribusi signifikan pada analisis, seperti “dan”, “di”, atau “yang”.
- Tokenization: Proses memecah teks menjadi unit-unit kecil seperti kata atau kalimat yang disebut token.
- Validasi Model: Proses pengujian performa model untuk memastikan keakuratan dan keandalan prediksi terhadap data baru.
- Word Embedding: Representasi teks dalam bentuk vektor numerik yang merepresentasikan hubungan semantik antar kata.

INTISARI

Analisis sentimen pada platform media sosial seperti Twitter merupakan salah satu cara efektif untuk memahami opini masyarakat terhadap berbagai isu atau topik. Penelitian ini bertujuan untuk mengklasifikasikan sentimen pada tweet berbahasa Indonesia menggunakan algoritma Naive Bayes sebagai model klasifikasi. Teknik undersampling digunakan untuk mengatasi ketidakseimbangan distribusi kelas dalam dataset, yang terdiri dari tiga kategori: positif, negatif, dan netral.

Proses penelitian dimulai dengan preprocessing data, termasuk tahap pembersihan teks, tokenisasi, dan stemming. Data kemudian dilabeli menggunakan metode clustering tak terawasi (unsupervised clustering), untuk mempermudah klasifikasi awal. Model dievaluasi menggunakan metrik akurasi, precision, recall, dan F1-score, serta dilengkapi dengan analisis distribusi data, visualisasi word cloud, dan confusion matrix untuk mengevaluasi performa lebih lanjut.

Hasil penelitian menunjukkan bahwa penerapan teknik undersampling berhasil meningkatkan keseimbangan distribusi kelas, sementara model Naive Bayes menunjukkan performa yang baik dalam mendeteksi sentimen pada tweet bahasa Indonesia. Nilai akurasi model mencapai 0.90, dengan precision 0.91, recall 0.90, dan F1-score 0.90. Analisis mendalam pada confusion matrix mengungkapkan bahwa model memiliki tingkat kesalahan yang rendah pada prediksi sentimen netral dibandingkan sentimen positif dan negatif.

Penelitian ini diharapkan dapat menjadi acuan untuk pengembangan sistem analisis sentimen yang lebih akurat dan efisien, serta memberikan kontribusi signifikan pada bidang Natural Language Processing dan analisis opini publik.

Kata kunci: Analisis Sentimen, Twitter, Naive Bayes, Ketidakseimbangan Kelas, Teknik Undersampling.

ABSTRACT

Sentiment analysis on social media platforms such as Twitter is an effective way to understand public opinion on various issues or topics. This study aims to classify sentiments in Indonesian-language tweets using the Naive Bayes algorithm as the classification model. The undersampling technique is applied to address class imbalance in the dataset, which consists of three categories: positive, negative, and neutral.

The research process begins with data preprocessing, including text cleaning, tokenization, and stemming. The data is then labeled using an unsupervised clustering method to facilitate initial classification. The model is evaluated using accuracy, precision, recall, and F1-score metrics, along with data distribution analysis, word cloud visualization, and confusion matrix analysis for further evaluation.

The results demonstrate that the application of undersampling effectively improved class distribution balance, while the Naive Bayes model showed good performance in detecting sentiments in Indonesian tweets. The model achieved an accuracy of 0.90, with precision 0.91, recall 0.90, and F1-score 0.90. Detailed analysis of the confusion matrix revealed that the model had a lower error rate in predicting neutral sentiments compared to positive and negative sentiments.

This study is expected to serve as a reference for developing more accurate and efficient sentiment analysis systems, contributing significantly to the field of Natural Language Processing and public opinion analysis.

Keywords: *Sentiment Analysis, Twitter, Naive Bayes, Class Imbalance, Undersampling Technique.*