

**ANALISIS SENTIMEN TWEET *TOXIC* MENGGUNAKAN  
ALGORITMA NAÏVE BAYES DAN SVM**

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**IRFAN ANDRYANTO**

**21.11.4398**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA**

**2025**

**ANALISIS SENTIMEN TWEET *TOXIC* MENGGUNAKAN  
ALGORITMA NAÏVE BAYES DAN SVM**

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**IRFAN ANDRYANTO**

**21.11.4398**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA  
2025**

## HALAMAN PERSETUJUAN



## HALAMAN PENGESAHAN

### SKRIPSI

#### HALAMAN PENGESAHAN SKRIPSI ANALISIS SENTIMEN TWEET *TOXIC* MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SVM

yang disusun dan diajukan oleh

**Irfan Andryanto**

**21.11.4398**

Telah dipertahankan di depan Dewan Penguji  
pada tanggal 12 Februari 2025

#### Susunan Dewan Penguji

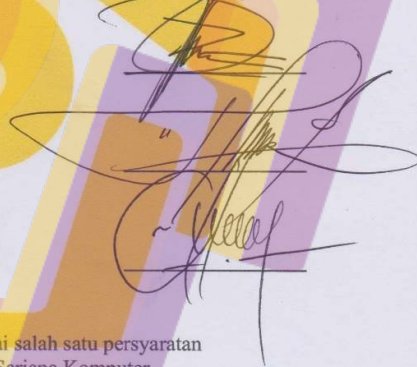
Nama Penguji

Arifiyanto Hadinegoro, S.Kom., M.T.  
NIK. 190302289

Haryoko, S.Kom., M.Cs.  
NIK. 190302286

Majid Rahardi, S.Kom., M.Eng.  
NIK. 190302393

Tanda Tangan



Skripsi ini telah diterima sebagai salah satu persyaratan  
untuk memperoleh gelar Sarjana Komputer  
Tanggal 12 Februari 2025

DEKAN FAKULTAS ILMU KOMPUTER



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

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

### HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Irfan Andryanto  
NIM : 21.11.4398

Menyatakan bahwa Skripsi dengan judul berikut:

**Analisis Sentimen Tweet Toxic Menggunakan Algoritma Naïve Bayes dan SVM**

Dosen Pembimbing : Majid Rahardi, S.Kom., M.Eng.

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, 12 Februari 2025

Yang Menyatakan,



Irfan Andryanto

## KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "*Analisis Sentimen Tweet Toxic Menggunakan Algoritma Naïve Bayes dan SVM*" dengan baik.

Penulisan skripsi tidak lepas dari liku-liku dan tantangan yang penulis hadapi. Namun berkat dukungan dan doa dari banyak pihak, penulis akhirnya dapat menyelesaikan tugas akhir ini. Oleh karena itu penulis mengucapkan terima kasih kepada :

1. Bapak Majid Rahardi, S.Kom., M.Eng., selaku pembimbing utama, atas arahan, motivasi, dan masukan yang sangat berarti selama proses penyusunan skripsi ini.
2. Orang tua dan keluarga tercinta atas doa, dukungan moral, dan materi yang senantiasa diberikan selama ini.
3. Sahabat dan rekan seperjuangan yang telah memberikan semangat dan bantuan selama proses penelitian dan penulisan skripsi ini.

Dalam penulisan ini, penulis menyadari masih adanya kekurangan. Oleh karena itu, dengan kerendahan hati penulis memohon maaf atas segala kesalahan baik disengaja maupun tidak disengaja.

Semoga skripsi ini dapat memberikan kontribusi positif dan menjadi referensi yang bermanfaat bagi pembaca dan penelitian selanjutnya.

Yogyakarta, 30 Desember 2024

Penulis

## DAFTAR ISI

|                                           |      |
|-------------------------------------------|------|
| HALAMAN PERSETUJUAN.....                  | ii   |
| HALAMAN PENGESAHAN .....                  | iii  |
| HALAMAN PERNYATAAN KEASLIAN SKRIPSI ..... | iv   |
| KATA PENGANTAR .....                      | v    |
| DAFTAR ISI.....                           | vi   |
| DAFTAR TABEL.....                         | viii |
| DAFTAR GAMBAR .....                       | ix   |
| DAFTAR LAMBANG DAN SINGKATAN .....        | xi   |
| DAFTAR ISTILAH .....                      | xii  |
| INTISARI .....                            | xiv  |
| <i>ABSTRACT</i> .....                     | xv   |
| BAB I PENDAHULUAN.....                    | 1    |
| 1.1 Latar Belakang .....                  | 1    |
| 1.2 Rumusan Masalah .....                 | 3    |
| 1.3 Batasan Masalah .....                 | 3    |
| 1.4 Tujuan Penelitian .....               | 4    |
| 1.5 Manfaat Penelitian .....              | 4    |
| 1.6 Sistematika Penulisan .....           | 4    |
| BAB II TINJAUAN PUSTAKA .....             | 7    |
| 2.1 Studi Literatur .....                 | 7    |
| 2.2 Dasar Teori.....                      | 14   |
| BAB III METODE PENELITIAN .....           | 19   |

|                                   |                                              |    |
|-----------------------------------|----------------------------------------------|----|
| 3.1                               | Alur Penelitian .....                        | 19 |
| 3.1.1                             | Data Akuisisi.....                           | 20 |
| 3.1.2                             | Data Preprocessing.....                      | 21 |
| A.                                | Proses Penerjemahan ke Bahasa Indonesia..... | 21 |
| B.                                | Duplicates Check .....                       | 23 |
| C.                                | Null Check .....                             | 24 |
| D.                                | Balance Check .....                          | 24 |
| E.                                | Data Cleaning.....                           | 27 |
| F.                                | Word Handle .....                            | 28 |
| G.                                | Stopword Removal dan Word Tokenization.....  | 30 |
| 3.1.3                             | Modeling .....                               | 31 |
| 3.1.4                             | Evaluasi.....                                | 33 |
| 3.2                               | Alat dan Bahan.....                          | 34 |
| 3.2.1                             | Alat.....                                    | 34 |
| 3.2.2                             | Bahan .....                                  | 35 |
| BAB IV HASIL DAN PEMBAHASAN ..... |                                              | 37 |
| 4.1                               | Pembersihan Data .....                       | 37 |
| 4.2                               | <i>Hyperparameter Tuning</i> .....           | 38 |
| 4.3                               | Countvectorizer TF-IDF .....                 | 39 |
| 4.4                               | Evaluasi <i>Confusion Matrix</i> .....       | 39 |
| BAB V PENUTUP .....               |                                              | 48 |
| 5.1                               | Kesimpulan .....                             | 48 |
| 5.2                               | Saran .....                                  | 48 |
| REFERENSI .....                   |                                              | 50 |



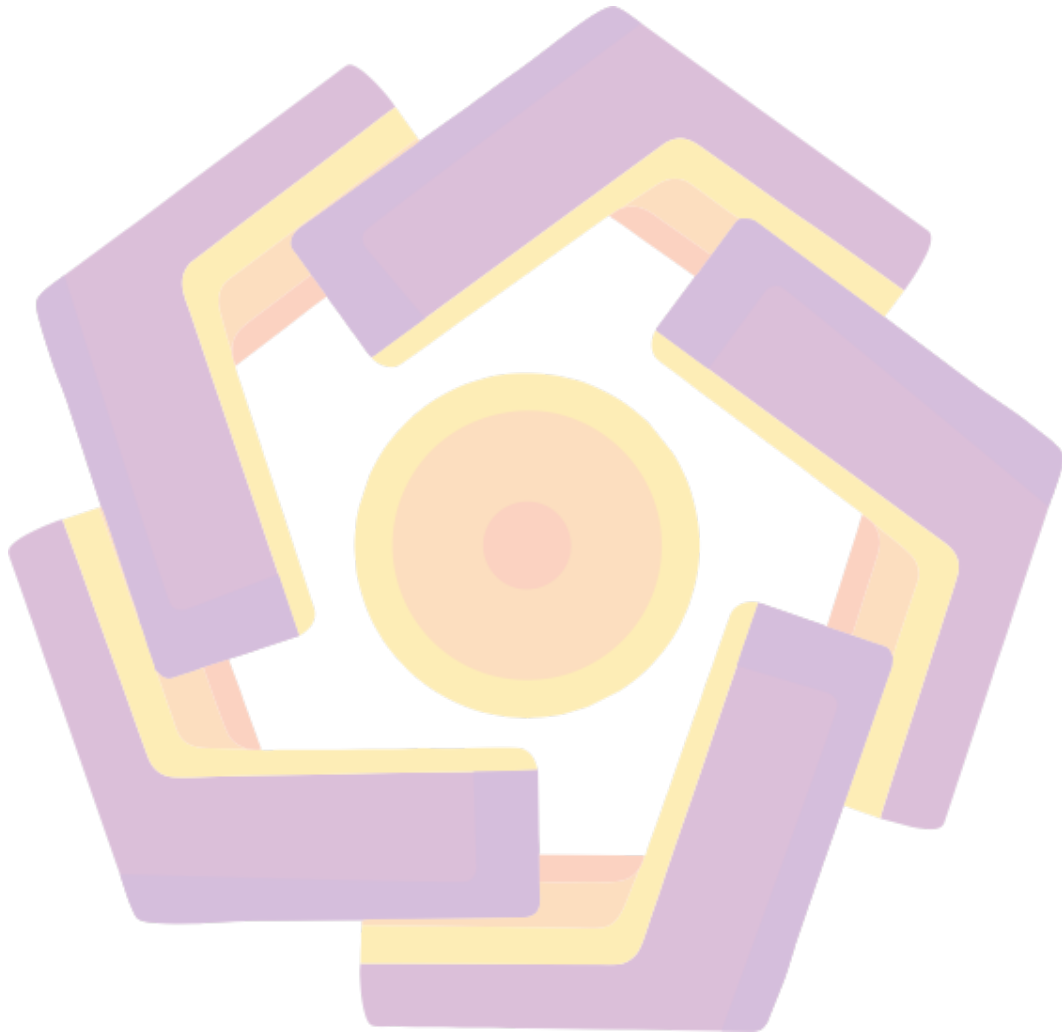
## DAFTAR TABEL

|                                                                           |    |
|---------------------------------------------------------------------------|----|
| Tabel 2.1 Keaslian Penelitian .....                                       | 10 |
| Tabel 3.1 Contoh data positif dengan label 0 .....                        | 20 |
| Tabel 3.2 Contoh data negatif dengan label 1 .....                        | 21 |
| Tabel 3.3 Proses Tokenisasi.....                                          | 30 |
| Tabel 3.4 Proses Stopword Removal .....                                   | 30 |
| Tabel 3.5 Parameter Model Naïve Bayes.....                                | 31 |
| Tabel 3.6 Parameter Model SVM .....                                       | 32 |
| Tabel 4.1 Data Sebelum Pembersihan Data.....                              | 37 |
| Tabel 4.2 Data Sesudah Pembersihan Data .....                             | 38 |
| Tabel 4.3 Parameter Hasil <i>Grid Search</i> .....                        | 38 |
| Tabel 4.4 Performa Model Naïve Bayes Inggris dengan Undersampling .....   | 40 |
| Tabel 4.5 Performa Model Naïve Bayes Indonesia dengan Undersampling ..... | 40 |
| Tabel 4.6 Performa Model SVM Inggris dengan Undersampling.....            | 41 |
| Tabel 4.7 Performa Model SVM Indonesia dengan Undersampling.....          | 41 |
| Tabel 4.8 Performa Model Naïve Bayes Inggris dengan SMOTE .....           | 43 |
| Tabel 4.9 Performa Model Naïve Bayes Indonesia dengan SMOTE .....         | 43 |
| Tabel 4.10 Performa Model SVM Inggris dengan SMOTE .....                  | 44 |
| Tabel 4.11 Performa Model SVM Indonesia dengan SMOTE.....                 | 44 |

## DAFTAR GAMBAR

|                                                                                                |    |
|------------------------------------------------------------------------------------------------|----|
| Gambar 3.1 Alur Penelitian .....                                                               | 19 |
| Gambar 3.2 Tampilan Dataset di Excel .....                                                     | 20 |
| Gambar 3.3 Dataset Asli Bahasa Inggris Tweet Non-Toxic.....                                    | 22 |
| Gambar 3.4 Dataset Asli Bahasa Inggris Tweet Toxic.....                                        | 22 |
| Gambar 3.5 Dataset Terjemahan Bahasa Indonesia Tweet Non-Toxic .....                           | 22 |
| Gambar 3.6 Dataset Terjemahan Bahasa Indonesia Tweet Toxic .....                               | 22 |
| Gambar 3.7 Hasil Kode Duplicate Check.....                                                     | 23 |
| Gambar 3.8 Hasil Kode NULL Check.....                                                          | 24 |
| Gambar 3.9 Keseimbangan Dataset Bahasa Inggris Sebelum Undersampling .....                     | 24 |
| Gambar 3.10 Keseimbangan Dataset Bahasa Indonesia Sebelum Undersampling .....                  | 25 |
| Gambar 3.11 Keseimbangan Data Bahasa Inggris Setelah Undersampling (RandomUndersampler).....   | 26 |
| Gambar 3.12 Keseimbangan Data Bahasa Indonesia Setelah Undersampling (RandomUndersampler)..... | 26 |
| Gambar 3.13 Keseimbangan Data Bahasa Inggris Setelah Oversampling (SMOTE) .....                | 27 |
| Gambar 3.14 Keseimbangan Data Bahasa Indonesia Setelah Oversampling (SMOTE) .....              | 27 |
| Gambar 3.15 Kode Untuk Pembersihan Data.....                                                   | 28 |
| Gambar 3.16 Potongan Kode Word Handle .....                                                    | 29 |
| Gambar 3.17 Potongan Kode Word Handle .....                                                    | 29 |
| Gambar 3.18 Tampilan Dataset di Website Kaggle.....                                            | 36 |
| Gambar 4.1 Kode Untuk Pembersihan Data.....                                                    | 37 |
| Gambar 4.2 Pembobotan Hasil TF-IDF.....                                                        | 39 |
| Gambar 4.3 Perbandingan Kinerja Model dengan Undersampling .....                               | 42 |
| Gambar 4.4 Perbandingan Kinerja Model dengan SMOTE .....                                       | 45 |
| Gambar 4.5 Perbandingan Kinerja Undersampling dan SMOTE Dataset Bahasa Inggris .....           | 46 |

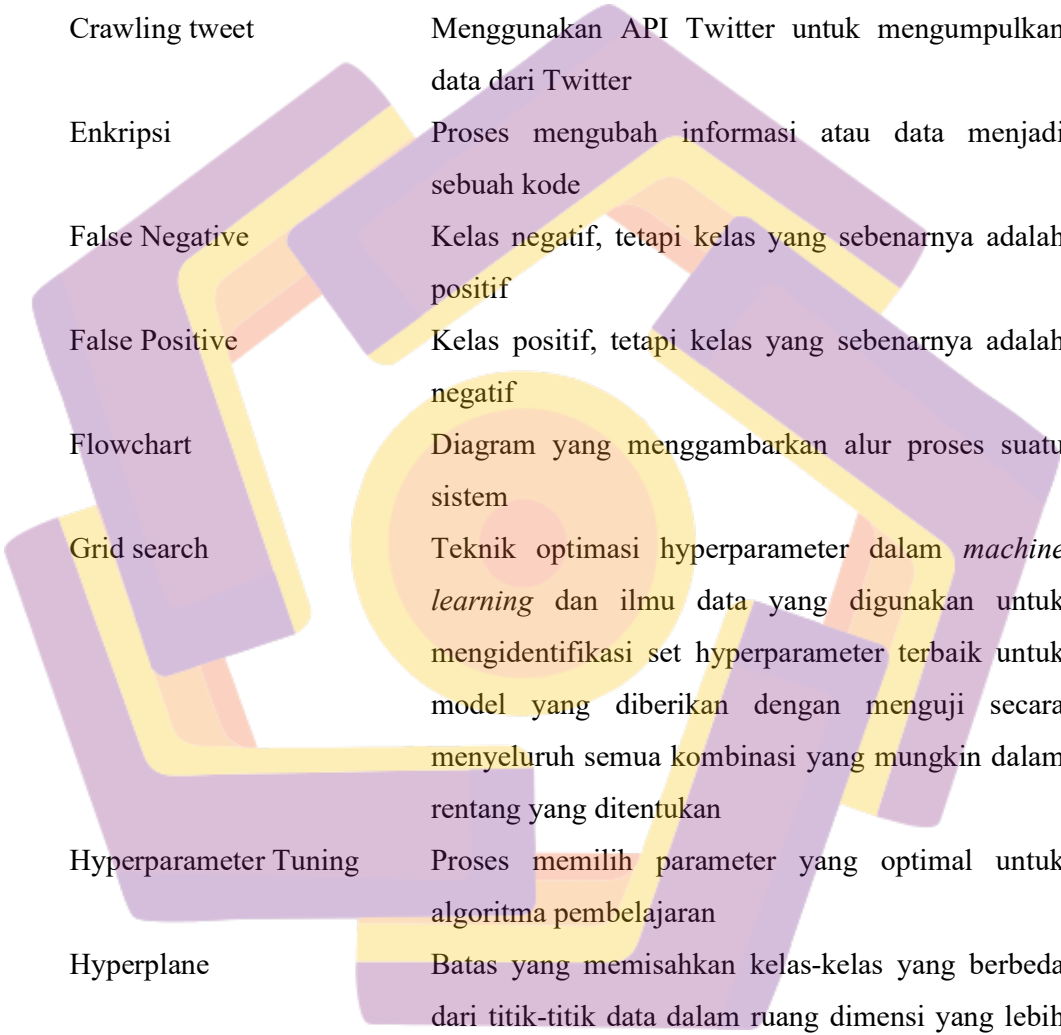
|                                                                                        |    |
|----------------------------------------------------------------------------------------|----|
| Gambar 4.6 Perbandingan Kinerja Undersampling dan SMOTE Dataset Bahasa Indonesia ..... | 46 |
|----------------------------------------------------------------------------------------|----|



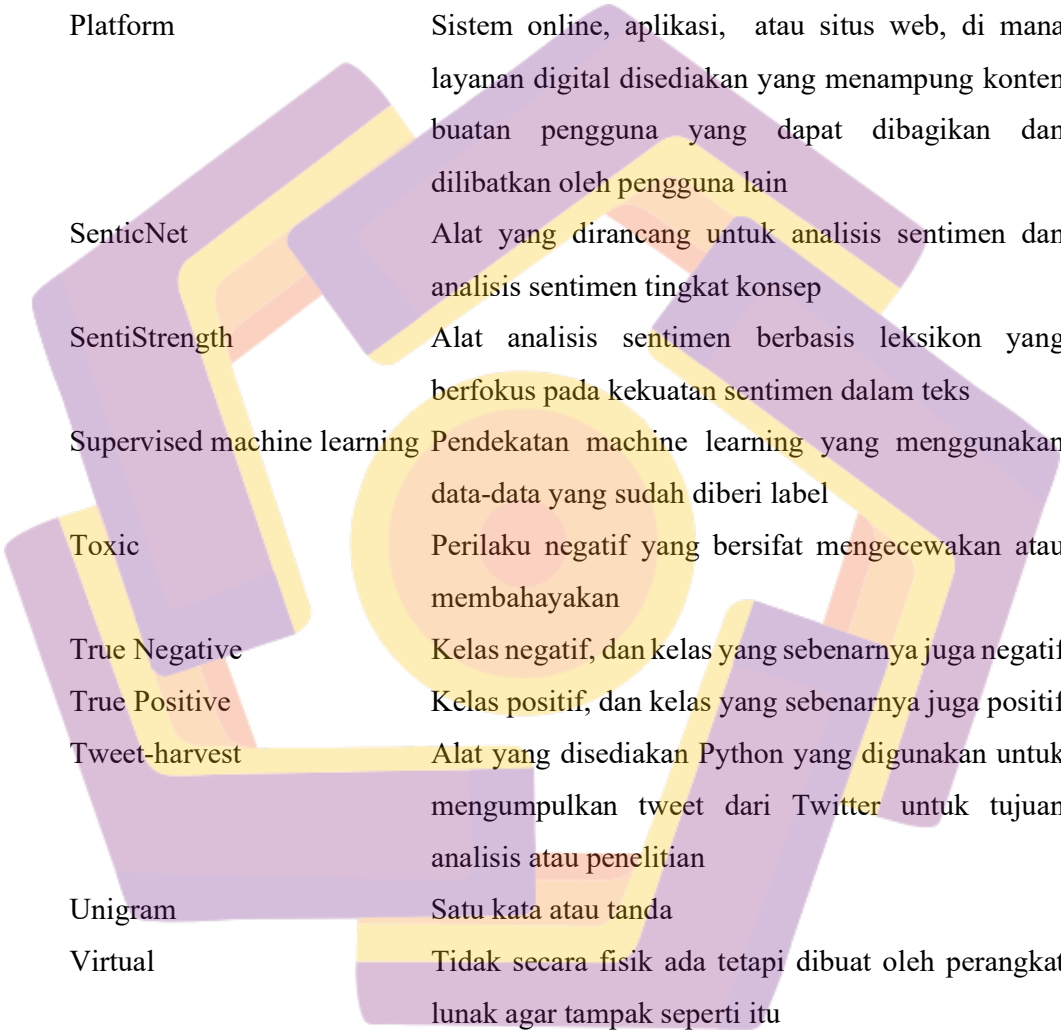
## DAFTAR LAMBANG DAN SINGKATAN

|                      |                                                                             |
|----------------------|-----------------------------------------------------------------------------|
| NB                   | Naïve Bayes                                                                 |
| SVM                  | Support Vector Machines                                                     |
| TF-IDF               | Term Frequency-Inverse Document Frequency                                   |
| BERT                 | Bidirectional encoder representations from transformers                     |
| RBF                  | Radial Basis Function                                                       |
| IBM                  | International Business Machines                                             |
| VADER                | Valence Aware Dictionary and sEntiment Reasoner                             |
| TF                   | Term Frequency                                                              |
| IDF                  | Inverse Document Frequency                                                  |
| $freq(t, d)$         | Jumlah kemunculan istilah $t$ dalam dokumen $d$                             |
| $ d $                | Jumlah total istilah dalam dokumen                                          |
| $ D $                | Jumlah dokumen/teks dalam sebuah korpus                                     |
| $ d \in D: t \in d $ | Jumlah dokumen/teks yang mengandung kata $t$                                |
| $P(C_i)$             | <i>Prior Probability</i> dari $X$                                           |
| $P(X)$               | Jumlah probabilitas <i>tuple</i> yang muncul                                |
| $P(C_i X)$           | <i>Posterior Probability</i> , probabilitas dari $C_i$ yang diberikan fakta |
| $P(X C_i)$           | Likelihood Probability, nilai parameter dengan kemungkinan<br>paling besar  |
| $f(x)$               | Fungsi prediksi                                                             |
| $w$                  | Vektor normal hyperplane                                                    |
| $x$                  | Vektor fitur input                                                          |
| $b$                  | Bias atau <i>intercept</i>                                                  |
| HTML                 | Hyper Text Markup Language                                                  |
| NLTK                 | Natural Language Toolkit                                                    |
| CPU                  | Central Processing Unit                                                     |
| RAM                  | Random Access Memory                                                        |
| GHz                  | Giga hertz                                                                  |
| GB                   | Giga Bytes                                                                  |

## DAFTAR ISTILAH



|                       |                                                                                                                                                                                                                                                                    |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Bias                  | Penyimpangan sistematis dalam hasil analisis atau prediksi yang dapat disebabkan oleh berbagai faktor.                                                                                                                                                             |
| Bigram                | Urutan dua kata atau tanda yang berurutan                                                                                                                                                                                                                          |
| Crawling tweet        | Menggunakan API Twitter untuk mengumpulkan data dari Twitter                                                                                                                                                                                                       |
| Enkripsi              | Proses mengubah informasi atau data menjadi sebuah kode                                                                                                                                                                                                            |
| False Negative        | Kelas negatif, tetapi kelas yang sebenarnya adalah positif                                                                                                                                                                                                         |
| False Positive        | Kelas positif, tetapi kelas yang sebenarnya adalah negatif                                                                                                                                                                                                         |
| Flowchart             | Diagram yang menggambarkan alur proses suatu sistem                                                                                                                                                                                                                |
| Grid search           | Teknik optimasi hyperparameter dalam <i>machine learning</i> dan ilmu data yang digunakan untuk mengidentifikasi set hyperparameter terbaik untuk model yang diberikan dengan menguji secara menyeluruh semua kombinasi yang mungkin dalam rentang yang ditentukan |
| Hyperparameter Tuning | Proses memilih parameter yang optimal untuk algoritma pembelajaran                                                                                                                                                                                                 |
| Hyperplane            | Batas yang memisahkan kelas-kelas yang berbeda dari titik-titik data dalam ruang dimensi yang lebih tinggi                                                                                                                                                         |
| Korpus                | Kumpulan teks tertulis                                                                                                                                                                                                                                             |
| Laplace Smoothing     | Teknik yang digunakan untuk menangani masalah probabilitas nol dalam model statistik                                                                                                                                                                               |



|                             |                                                                                                                                                                          |
|-----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Machine Learning            | Mesin yang dikembangkan untuk melakukan suatu tugas tertentu berdasarkan data yang telah diberikan dan terus berkembang dengan sendirinya                                |
| Multilingual                | Menggunakan beberapa bahasa                                                                                                                                              |
| NULL                        | Tidak memiliki nilai                                                                                                                                                     |
| Platform                    | Sistem online, aplikasi, atau situs web, di mana layanan digital disediakan yang menampung konten buatan pengguna yang dapat dibagikan dan dilibatkan oleh pengguna lain |
| SenticNet                   | Alat yang dirancang untuk analisis sentimen dan analisis sentimen tingkat konsep                                                                                         |
| SentiStrength               | Alat analisis sentimen berbasis leksikon yang berfokus pada kekuatan sentimen dalam teks                                                                                 |
| Supervised machine learning | Pendekatan machine learning yang menggunakan data-data yang sudah diberi label                                                                                           |
| Toxic                       | Perilaku negatif yang bersifat mengecewakan atau membahayakan                                                                                                            |
| True Negative               | Kelas negatif, dan kelas yang sebenarnya juga negatif                                                                                                                    |
| True Positive               | Kelas positif, dan kelas yang sebenarnya juga positif                                                                                                                    |
| Tweet-harvest               | Alat yang disediakan Python yang digunakan untuk mengumpulkan tweet dari Twitter untuk tujuan analisis atau penelitian                                                   |
| Unigram                     | Satu kata atau tanda                                                                                                                                                     |
| Virtual                     | Tidak secara fisik ada tetapi dibuat oleh perangkat lunak agar tampak seperti itu                                                                                        |

## INTISARI

Penggunaan media sosial Twitter/X yang terus meningkat menyebabkan persebaran konten yang lebih bervariasi. Sayangnya tidak semua konten yang tersebar bersifat positif, ada beberapa konten yang bersifat negatif salah satunya adalah konten yang bersifat toxic. Konten toxic ini dapat berdampak negatif pada kenyamanan dan kesehatan mental pengguna. Oleh karena itu, penting untuk mengembangkan metode yang efektif dalam mendeteksi dan menganalisis sentimen pada tweet toxic. Penelitian ini bertujuan untuk melakukan analisis sentimen pada tweet toxic menggunakan algoritma Naïve Bayes dan Support Vector Machine (SVM). Dataset yang digunakan merupakan dataset berisi tweet toxic dan non-toxic yang didapat dari website Kaggle yang selanjutnya di translate menggunakan Google Sheet untuk mendapatkan dataset berbahasa Indonesia. Dalam proses analisis, data teks diolah menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency) untuk merepresentasikan teks dalam bentuk numerik. Selanjutnya, model Naïve Bayes dan SVM dilatih dan dievaluasi menggunakan metode Confusion Matrix. Hasil penelitian menunjukkan bahwa algoritma SVM memiliki performa yang lebih baik dibandingkan Naïve Bayes dalam hal akurasi, presisi, dan recall pada dataset yang digunakan. Model Naïve Bayes mendapatkan rata-rata akurasi 86.67% untuk Bahasa Inggris dan 84.26% untuk Bahasa Indonesia. Sedangkan untuk model SVM mendapatkan rata-rata akurasi sebesar 90.52% untuk Bahasa Inggris dan 88.09% untuk Bahasa Indonesia. Dengan demikian, penelitian ini membuktikan bahwa SVM memiliki kinerja yang lebih unggul dalam mendeteksi konten toxic multilingual di media sosial.

**Kata kunci:** Analisis Sentimen, Naïve Bayes, SVM, Twitter, Toxic.

## **ABSTRACT**

*The increasing use of Twitter/X social media has led to more varied content distribution. Unfortunately, not all content that is spread is positive, there is some content that is negative, one of which is toxic content. This toxic content can have a negative impact on user comfort and mental health. Therefore, it is important to develop an effective method for detecting and analysing sentiment in toxic tweets. This research aims to perform sentiment analysis on toxic tweets using Naïve Bayes algorithm and Support Vector Machine (SVM). The dataset used is a dataset containing toxic and non-toxic tweets obtained from the Kaggle website which is then translated using Google Sheet to get an Indonesian language dataset. In the analysis process, the text data is processed using the TF-IDF (Term Frequency-Inverse Document Frequency) method to represent the text in numerical form. Furthermore, Naïve Bayes and SVM models are trained and evaluated using K-Fold cross validation technique. The results showed that the SVM algorithm performed better than Naïve Bayes in terms of accuracy, precision, and recall on the datasets used. The Naïve Bayes model gets an average accuracy of 86.67% for English and 84.26% for Indonesian. Meanwhile, the SVM model obtained an average accuracy of 90.52% for English and 88.09% for Indonesian. Thus, this research proves that SVM has superior performance in detecting toxic multilingual content on social media.*

**Keyword:** Sentiment Analysis, Naïve Bayes, SVM, Twitter, Toxic