**ANALISIS SENTIMEN ULASAN PRODUK *E-COMMERCE* MENGGUNAKAN *MACHINE LEARNING (STUDI KASUS: AMATERASUN SUNSCREN SERUM)***

**TUGAS BESAR DATA MINING**

****

**Oleh:**

**M. Fachriza Farhan (714220005)**

**Fulandi Hudza Grahitama (714220010)**

**Muhammad Rafli Alfarisi (714220008)**

**Fathya Fathimah Azzahra (714220012)**

**DIPLOMA IV TEKNIK INFORMATIKA**

**SEKOLAH VOKASI**

**UNIVERSITAS LOGISTIK & BISNIS INTERNASIONAL**

**BANDUNG**

# HALAMAN PERNYATAAN ORISINALITAS

Laporan tugas besar ini adalah hasil karya kami sendiri, dan semua sumber baik yang dikutip maupun dirujuk telah kami nyatakan dengan benar. Bilamana di kemudian hari ditemukan bahwa karya tulis ini menyalahi peraturan yang ada berkaitan dengan etika dan kaidah penulisan karya ilmiah yang berlaku, maka kami bersedia dituntut dan diproses sesuai dengan ketentuan yang berlaku.

Yang menyatakan,

Nama : M. Fachriza Farhan  
NIM : 714220005

Tanda Tangan: ...............................

Tanggal: 10 Juli 2025

Mengetahui,

Ketua :.................................. (.......tanda tangan. .....)

Dosen Pengampu Mata Kuliah Data Mining : .................................. (.......tanda tangan. .....)

# KATA PENGANTAR

Puji syukur kami panjatkan kepada Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga kami dapat menyelesaikan laporan Tugas Besar Data Mining ini yang berjudul *"Analisis Sentimen Ulasan Produk E-Commerce Menggunakan Algoritma Machine Learning (Studi Kasus: Amaterasun Sunscreen Serum)"*.

Laporan ini disusun untuk memenuhi tugas akhir mata kuliah Data Mining pada Program Studi D4 Teknik Informatika, Universitas Logistik dan Bisnis Internasional.

Kami mengucapkan terima kasih kepada:

* Dosen pengampu mata kuliah Data Mining, atas bimbingan dan ilmunya selama perkuliahan berlangsung.
* Orang tua dan keluarga yang selalu memberikan dukungan moril dan semangat.
* Rekan satu kelompok atas kerja sama dan komitmen dalam menyelesaikan tugas ini bersama.

Kami menyadari bahwa laporan ini masih memiliki kekurangan. Oleh karena itu, kritik dan saran yang membangun sangat kami harapkan demi perbaikan di masa mendatang.

Bandung,10 Juli 2025

Penyusun,

M. Fachriza Farhan

Fulandi Hudza Grahitama

Muhammad Rafli Alfarissi

Fathya Fathimah Azahra

# HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Universitas Logistik dan Bisnis Internasional, saya yang bertanda tangan di bawah ini:

Nama: M. Fachriza Farhan  
NIM : 714220005

Selaku Ketua Kelompok, menyatakan menyetujui untuk memberikan kepada Universitas Logistik dan Bisnis Internasional, hak bebas royalti noneksklusif (non-exclusive royalty free right) atas karya ilmiah kami yang berjudul:

"Analisis Sentimen Ulasan Produk E-Commerce Menggunakan Algoritma Machine Learning (Studi Kasus: Amaterasun Sunscreen Serum)"

Beserta perangkat yang ada (jika diperlukan). Dengan hak ini, ULBI berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan tugas akhir kami selama tetap mencantumkan nama kami sebagai penulis/pemilik hak cipta.

Demikian pernyataan ini saya buat dengan sebenar-benarnya.

Dibuat di : Bandung  
Pada tanggal : 10 Juli 2025

Yang menyatakan,

M. Fachriza Farhan  
Ketua Kelompok

# ABSTRAK

Analisis sentimen merupakan metode penting untuk mengevaluasi opini konsumen terhadap suatu produk, khususnya di *platform e-commerce*. Penelitian ini bertujuan untuk mengklasifikasikan sentimen ulasan pengguna terhadap produk *Amaterasun Sunscreen Serum* ke dalam kategori positif dan negatif menggunakan pendekatan machine learning. Data ulasan diperoleh dari platform X (Twitter) dan website *Female Daily*.

Setelah dilakukan *preprocessing* dan ekstraksi fitur menggunakan TF-IDF, dilakukan pelatihan model menggunakan tiga algoritma: *Logistic Regression* (LR), *Support Vector Machine* (SVM), dan *Naïve Bayes* (NB). Hasil evaluasi menunjukkan bahwa model SVM memiliki performa terbaik dalam hal akurasi klasifikasi.

Penelitian ini memberikan kontribusi terhadap pengembangan sistem pendukung keputusan dan pemahaman terhadap persepsi konsumen dalam industri kecantikan berbasis data.

Kata Kunci: *analisis sentimen, e-commerce, machine learning, Logistic Regression, Support Vector Machine, naïve bayes, Amaterasun, TF-IDF.*

# ABSTRACT

*Sentiment analysis is a crucial technique for evaluating consumer opinions about a product, especially on e-commerce platforms. This study aims to classify user reviews of the Amaterasun Sunscreen Serum into positive and negative categories using machine learning approaches. The review data were collected from X (Twitter) and the Female Daily website.*

*After preprocessing and feature extraction using TF-IDF, model training was conducted using three algorithms: Logistic Regression (LR), Support Vector Machine (SVM), and Naïve Bayes (NB). The evaluation results show that the SVM model achieved the best performance in terms of classification accuracy.*

*This research contributes to the development of decision support systems and understanding consumer perception in the beauty industry through data-driven approaches.*

*Keywords: sentiment analysis, e-commerce, machine learning, Logistic Regression, Support Vector Machine, naïve bayes, Amaterasun, TF-IDF.*

# DAFTAR ISI

[HALAMAN PERNYATAAN ORISINALITAS 2](#_Toc203920044)

[KATA PENGANTAR 3](#_Toc203920045)

[HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS 4](#_Toc203920046)

[ABSTRAK 5](#_Toc203920047)

[ABSTRACT 6](#_Toc203920048)

[DAFTAR ISI 7](#_Toc203920049)

[DAFTAR TABEL 9](#_Toc203920050)

[DAFTAR GAMBAR 10](#_Toc203920051)

[DAFTAR RUMUS 11](#_Toc203920052)

[DAFTAR NOTASI 12](#_Toc203920053)

[BAB I PENDAHULUAN 13](#_Toc203920054)

[1.1 Latar Belakang 13](#_Toc203920055)

[1.2 Rumusan Masalah 13](#_Toc203920056)

[1.3 Tujuan Penelitian 13](#_Toc203920057)

[1.4 Manfaat Penelitian 14](#_Toc203920058)

[1.5 Ruang Lingkup 14](#_Toc203920059)

[BAB II TINJAUAN PUSTAKA 15](#_Toc203920060)

[2.1 Landasan Teori dan Konsep 15](#_Toc203920063)

[2.1.1 Analisis Sentimen 15](#_Toc203920064)

[2.1.2 *Natural Language Processing* (NLP) 15](#_Toc203920065)

[2.1.3 Algoritma Klasifikasi 16](#_Toc203920066)

[2.1.4 Studi Terkait 17](#_Toc203920067)

[2.2 Diagram Alur Konsep Metodologi 19](#_Toc203920068)

[2.3 State of The Art 19](#_Toc203920069)

[BAB III METODOLOGI 21](#_Toc203920070)

[3.1 Tahapan Penelitian - Metodologi Data Mining (CRISP-DM) 21](#_Toc203920071)

[3.2 Deskripsi *Dataset* 22](#_Toc203920072)

[3.3 Algoritma dan Tools yang Digunakan 22](#_Toc203920073)

[3.4 Evaluasi Kinerja Model 23](#_Toc203920074)

[BAB IV HASIL DAN PEMBAHASAN 24](#_Toc203920075)

[4.1 Visualisasi dan EDA 24](#_Toc203920076)

[4.1.1 Distribusi Label Sentimen 24](#_Toc203920077)

[4.1.2 WordCloud 25](#_Toc203920078)

[4.1.3 Statistik Deskriptif 25](#_Toc203920079)

[4.2 Hasil *Preprocessing* dan Feature Engineering 26](#_Toc203920080)

[4.3 Hasil *Preprocessing* dan Pemodelan 26](#_Toc203920081)

[4.3.1 Confussion Matrix 27](#_Toc203920082)

[4.4 Interpretasi Hasil dan Analisis Komparatif 29](#_Toc203920083)

[4.5 Analisis Keunggulan dan Keterbatasan 30](#_Toc203920084)

[BAB V KESIMPULAN DAN SARAN 31](#_Toc203920085)

[5.1 Kesimpulan 31](#_Toc203920086)

[5.2 Jawaban atas Rumusan Masalah 31](#_Toc203920087)

[5.3 Saran 32](#_Toc203920088)

[DAFTAR PUSTAKA 33](#_Toc203920089)

[LAMPIRAN 35](#_Toc203920090)

# DAFTAR TABEL

[Tabel 2. 1 Tabel Studi Terkait 17](#_Toc203920092)

[Tabel 3. 1 Tabel Evaluasi Kinerja Model 22](#_Toc203067699)

[Tabel 4. 1 Hasil Evaluasi Performa Model Klasifikasi Sentimen…………………………….26](#_Toc203920329)

# DAFTAR GAMBAR

[Gambar 2. 1 Gambar Diagram Alur Konsep Metodologi 19](#_Toc203920210)

[Gambar 4. 1 Distribusi Label Sentimen………………………………………………………24](#_Toc203920245)

[Gambar 4. 2 WordCloud Positif 25](#_Toc203920246)

[Gambar 4. 3 WordCloud negatif 25](#_Toc203920247)

[Gambar 4. 4 Confussion Matrix - *Logistic Regression* 27](#_Toc203920248)

[Gambar 4. 5 Confussion Matrix - *Naive Bayes* 28](#_Toc203920249)

[Gambar 4. 6 Confussion Matrix – SVM 29](#_Toc203920250)

[Gambar. 1 Gambar *Dataset* Mentah atau Raw 35](#_Toc203920212)

# DAFTAR RUMUS

# DAFTAR NOTASI

# BAB I PENDAHULUAN

## Latar Belakang

Cara konsumen memberikan dan mendapatkan informasi mengenai produk telah berubah karena perkembangan teknologi digital dan *e-commerce*. Salah satu informasi penting yang mudah diakses adalah ulasan konsumen, yang menggambarkan opini atau sentimen terhadap sebuah produk, seperti produk *skincare* dan *sunscreen*. Dengan semakin banyaknya jumlah ulasan secara *online*, muncul kebutuhan untuk melakukan analisis sentiment secara otomatis menggunakan pendekatan *Natural Languange Processing (NLP).*

Analisis sentimen adalah teknik untuk mengidentifikasi opini subjektif dari teks dan mengkategorikannya menjadi positif, negatif, atau netral [1]. Dalam proyek ini, teknik *machine learning* modern seperti *Transformer* dari *Huggingface* dipakai untuk memberi label sentiment pada data, setelah itu model *machine learning* klasik seperti *Logistic Regression, Naïve Bayes,* dan *Support Vector Machine (SVM)* digunakan untuk membangun model klasifikasi sentiment.

Penelitian terdahulu menunjukkan bahwa algoritma seperti *LR, NB*, dan *SVM* sudah digunakan secara luas dalam berbagai bidang seperti ulasan *e-commerce*, film, dan aplikasi lain, dan memiliki performa yang beragam tergantung pada *dataset* dan *preprocessing* yang digunakan Haroon et al. dan Aziz et al [2], [3]. Pendekatan berbasis *Transformer* seperti *BERT* juga terbukti memberikan akurasi yang lebih tinggi dibanding metode klasik karena kemampuannya dalam memahami konteks kalimat[4].

## Rumusan Masalah

1. Bagaimana memanfaatkan model *Transformer* untuk memberi label sentimen pada ulasan produk *sunscreen*?
2. Sejauh mana algoritma *LR*, *NB*, dan *SVM* dapat memodelkan hasil labeling sentimen tersebut secara akurat?
3. Algoritma mana yang memberikan performa terbaik dalam klasifikasi sentimen ulasan produk?

## Tujuan Penelitian

1. Membangun *dataset* sentimen dari ulasan produk sunscreen dengan bantuan model Transformer.
2. Mengembangkan model klasifikasi sentimen menggunakan *LR*, *NB*, dan *SVM*. Membandingkan performa ketiga algoritma dan menentukan yang paling optimal dalam konteks analisis sentimen produk *sunscreen*.

## Manfaat Penelitian

Penelitian ini memberikan kontribusi dalam beberapa aspek penting. Dari sisi teknis, penelitian ini menunjukkan bagaimana kombinasi antara model *deep learning* modern seperti Transformer (misalnya BERT) dan algoritma *machine learning* klasik seperti *Naive Bayes*, *Logistic Regression*, serta *Support Vector Machine* dapat diterapkan secara efektif untuk tugas klasifikasi sentimen. Hal ini berguna bagi pengembang sistem analisis teks yang ingin mengintegrasikan kemampuan pemahaman konteks dari Transformer dengan efisiensi model pembelajaran klasik. Dari sisi praktis, hasil penelitian ini dapat dimanfaatkan oleh pelaku industri kecantikan dan *e-commerce* untuk memahami opini konsumen terhadap produk sunscreen secara otomatis, sehingga mendukung pengambilan keputusan dalam pengembangan produk, strategi pemasaran, dan pelayanan pelanggan. Selain itu, dari sisi akademis, penelitian ini dapat menjadi rujukan bagi peneliti lain yang tertarik pada pengembangan sistem klasifikasi teks berbasis NLP dan pada eksplorasi kombinasi pendekatan modern dan tradisional dalam analisis sentimen.

## Ruang Lingkup

Agar penelitian ini terfokus dan terarah, ruang lingkupnya dibatasi sebagai berikut:

1. **Jenis data:** Data yang digunakan berupa ulasan teks dari konsumen terhadap produk *sunscreen* yang diambil dari *platform e-commerce* atau forum daring.
2. **Labeling data:** Label sentimen (positif, negatif, netral) diperoleh menggunakan model *Transformer* dari *library* *Huggingface* (misalnya: BERT).
3. **Pemodelan klasifikasi:** Model yang digunakan untuk klasifikasi adalah *Naive Bayes*, *Logistic Regression*, dan *Support Vector Machine*. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, dan *F1-Score*.
4. **Bahasa:** Ulasan yang dianalisis terbatas pada teks berbahasa Indonesia.
5. ***Preprocessing*:** Melibatkan proses dasar seperti tokenisasi, penghapusan *stopwords*, dan *stemming*, disesuaikan dengan bahasa dan format teks ulasan.

# BAB II TINJAUAN PUSTAKA

Bab ini menyajikan landasan teoretis yang mendasari penelitian, mencakup konsep-konsep kunci dalam analisis sentimen, *machine learning*, dan pemrosesan teks. Selain itu, bab ini juga mengulas penelitian-penelitian terdahulu yang relevan (*studi terkait*) untuk memetakan posisi dan kebaruan (*state of the art*) dari penelitian ini, serta menyajikan kerangka pikir metodologi yang akan digunakan.



## Landasan Teori dan Konsep

Bagian ini menjelaskan secara mendalam mengenai teori, definisi, dan teknik yang menjadi fondasi dalam penelitian ini.

### Analisis Sentimen

Analisis sentimen merupakan proses yang bertujuan untuk mengidentifikasi, mengklasifikasikan, dan mengevaluasi opini atau perasaan seseorang terhadap suatu objek, topik, atau isu tertentu yang diekspresikan melalui teks. Biasanya, analisis ini diterapkan pada data dari media sosial, forum diskusi, atau ulasan daring. Tujuan utamanya adalah untuk mengetahui apakah opini yang disampaikan bersifat positif, negatif, atau netral, sehingga dapat memberikan gambaran umum mengenai persepsi publik terhadap suatu hal.

Dalam praktiknya, analisis sentimen dilakukan dengan memanfaatkan teknik text mining dan algoritma pembelajaran mesin seperti Naïve Bayes, *Logistic Regression*, *Support Vector Machine*, Random Forest, maupun model berbasis deep learning seperti BERT. Proses ini melibatkan beberapa tahapan, mulai dari pengumpulan data, pra-pemrosesan teks untuk menghilangkan unsur yang tidak relevan, pelabelan sentimen pada data, pembangunan model untuk klasifikasi, hingga evaluasi akurasi model dalam mengklasifikasikan sentimen [5], [6], [7].

### *Natural Language Processing* (NLP)

*Natural Language Processing* (NLP) adalah cabang dari kecerdasan buatan (AI) yang berfokus pada kemampuan komputer untuk memahami, menafsirkan, menganalisis, dan menghasilkan bahasa manusia secara otomatis. Teknologi ini memungkinkan mesin untuk berinteraksi dengan manusia menggunakan bahasa alami, baik dalam bentuk teks maupun suara, sehingga peran NLP menjadi sangat penting dalam menjembatani komunikasi antara manusia dan komputer.

NLP mencakup berbagai aplikasi, seperti analisis sentimen, pengenalan suara, penerjemahan otomatis, rangkuman teks otomatis, serta *chatbot* dan asisten virtual. Penerapannya sangat luas, mulai dari bidang pendidikan, kesehatan, bisnis, hingga industri penerbangan dan pengembangan *smart city*. Untuk mendukung kemampuannya, NLP memanfaatkan teknik *machine learning* guna mengklasifikasikan, mengekstrak, dan memahami informasi dari data berbasis bahasa manusia, sehingga proses pengolahan informasi menjadi lebih efisien dan akurat [8], [9], [10].

### Algoritma Klasifikasi

Penelitian ini akan menggunakan teknik klasifikasi dalam analisis sentimen sebagai metode untuk mengelompokkan teks ulasan, komentar, atau opini ke dalam kategori sentimen tertentu, seperti positif, negatif, atau netral. Proses klasifikasi ini dilakukan berdasarkan isi dan konteks bahasa yang digunakan dalam teks, sehingga dapat mencerminkan sikap atau perasaan pengguna terhadap suatu produk, layanan, atau isu tertentu.

Dengan memanfaatkan algoritma komputer, klasifikasi sentimen dilakukan secara otomatis untuk meningkatkan efisiensi dan akurasi dalam menganalisis data dalam skala besar. Pendekatan ini memungkinkan sistem untuk mengenali pola bahasa yang mencerminkan emosi atau opini, sehingga hasil yang diperoleh dapat digunakan sebagai dasar pengambilan keputusan dalam berbagai bidang, seperti pemasaran, pelayanan pelanggan, maupun pengembangan produk [11].

#### Naive Bayes Classifier (NBC)

*Naïve Bayes Classifier* (NBC) akan digunakan dalam penelitian ini sebagai salah satu algoritma klasifikasi dalam bidang *machine learning* yang didasarkan pada teorema Bayes. Algoritma ini mengasumsikan bahwa setiap fitur dalam data bersifat saling independen atau tidak saling bergantung, meskipun dalam praktiknya asumsi ini seringkali tidak sepenuhnya sesuai dengan kondisi sebenarnya.

Meskipun memiliki asumsi yang sederhana, NBC dikenal karena kemudahannya dalam implementasi serta efisiensinya dalam memproses data dalam jumlah besar. Selain itu, algoritma ini juga mampu memberikan performa yang cukup baik dalam berbagai tugas klasifikasi, termasuk analisis sentimen, sehingga menjadi pilihan yang relevan dan efektif dalam konteks penelitian ini [12].

#### Support Vector Machine (SVM)

*Support Vector Machine* (SVM) akan digunakan dalam penelitian ini sebagai salah satu algoritma machine learning yang efektif untuk tugas klasifikasi. Prinsip utama dari SVM adalah mencari hyperplane terbaik yang dapat memisahkan data ke dalam kelas-kelas yang berbeda dengan margin maksimum. Dengan kata lain, SVM berusaha memaksimalkan jarak antara *hyperplane* pemisah dan titik data terdekat dari masing-masing kelas, yang dikenal sebagai *support vectors*.

Keunggulan SVM terletak pada kemampuannya dalam menangani data berdimensi tinggi dan memberikan hasil klasifikasi yang akurat, terutama ketika batas antar kelas tidak dapat dipisahkan secara linear. Oleh karena itu, dalam konteks analisis sentimen, SVM menjadi salah satu pilihan algoritma yang tepat untuk mengelompokkan opini berdasarkan sentimen secara optimal [13].

#### Logistic Regresion

Regresi logistik akan digunakan dalam penelitian ini sebagai salah satu metode statistik untuk memodelkan hubungan antara satu atau lebih variabel independen (fitur) dengan variabel dependen yang bersifat biner, seperti sentimen positif atau negatif. Berbeda dengan regresi linier yang menghasilkan prediksi berupa nilai numerik, regresi logistik memprediksi probabilitas terjadinya suatu kejadian menggunakan fungsi logistik atau sigmoid.

Melalui pendekatan ini, hasil prediksi akan berada dalam rentang antara 0 hingga 1, yang kemudian diklasifikasikan ke dalam kelas tertentu berdasarkan ambang batas tertentu (threshold). Keunggulan regresi logistik terletak pada kesederhanaannya, interpretasi model yang jelas, serta kemampuannya dalam menangani masalah klasifikasi biner secara efektif, sehingga relevan untuk digunakan dalam analisis sentimen pada penelitian ini [5], [6], [7], [8], [9], [10], [11], [12], [13]

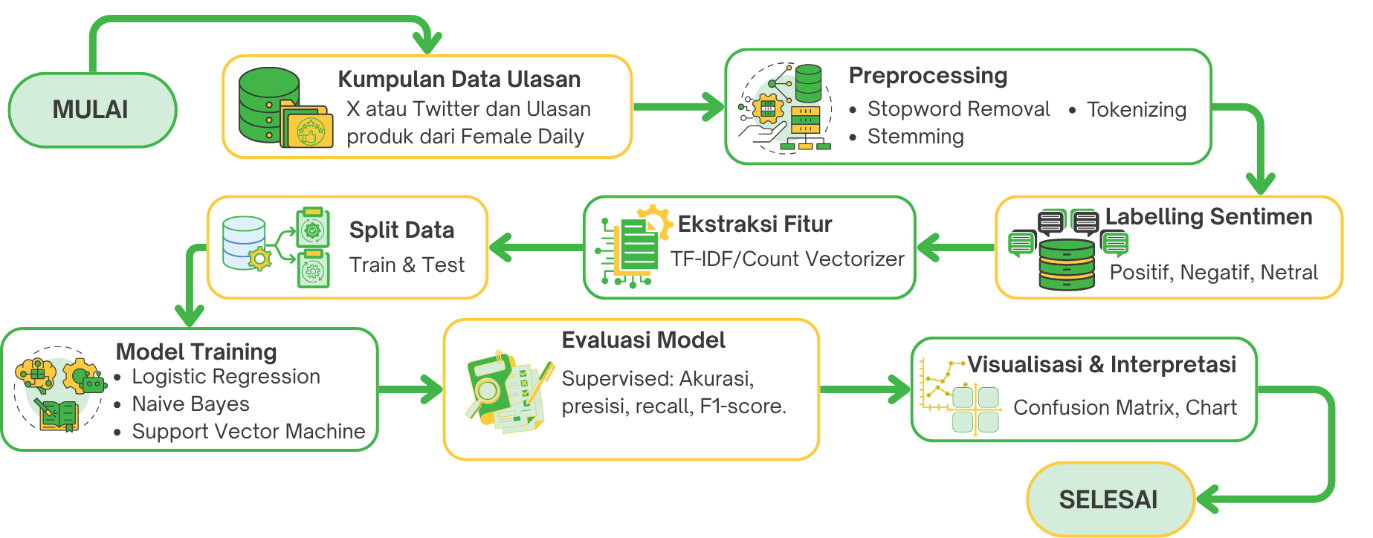
### Studi Terkait

Berikut adalah tinjauan beberapa penelitian relevan yang menjadi acuan dalam penelitian ini:

Tabel 2. Tabel Studi Terkait

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Penelitian (Tahun) | Judul Penelitian | Metode | Hasil Utama |
| 1 | Chenglerayen & Raja (2024) | Analisis ulasan Amazon untuk evolusi produk. | Transformer (BERT) | Menegaskan bahwa model Transformer seperti BERT menawarkan pemahaman kontekstual yang mendalam dan telah menjadi standar baru dalam NLP. |
| 2 | Aziz et al. (2023) | Perbandingan metode untuk analisis ulasan film. | Naïve Bayes, *Logistic Regression*, SVM | Melakukan perbandingan langsung antara tiga algoritma klasik, memberikan dasar untuk pemilihan model dalam penelitian ini. |
| 3 | Sharma et al. (2023) | Analisis sentimen pada ulasan film. | Naïve Bayes | Membuktikan bahwa Naïve Bayes memiliki performa yang kompetitif, serta unggul dalam kecepatan dan kesederhanaan. |
| 4 | Tan & Chow (2021) | Analisis sentimen pada ulasan game. | SVM, Machine Learning (umum) | Menunjukkan bahwa SVM terbukti efektif untuk data teks yang memiliki dimensi tinggi, yang relevan untuk data ulasan. |
| 5 | Anees et al. (2019) | Analisis performa klasifikasi dengan skema pembobotan istilah. | *Logistic Regression*, Classifier (umum) | Menemukan bahwa *Logistic Regression* memiliki keunggulan dalam hal akurasi dan kemampuan generalisasi di berbagai skenario. |

## Diagram Alur Konsep Metodologi



Gambar 2. Gambar Diagram Alur Konsep Metodologi

## State of The Art

Penelitian di bidang analisis sentimen telah mengalami evolusi yang pesat, bergerak dari pendekatan leksikon sederhana menuju model machine learning yang canggih. Pada awalnya, algoritma klasik seperti *Naïve Bayes* (NB), *Logistic Regression* (LR), dan *Support Vector Machine* (SVM) menjadi andalan utama. Ketiga model ini masih sangat relevan dan terbukti memiliki keunggulan masing-masing; *Naïve Bayes* dikenal karena kecepatan dan kesederhanaannya [5], *Logistic Regression* unggul dalam akurasi dan generalisasi [6], sementara SVM sangat efektif untuk menangani data teks berdimensi tinggi [7].

Dalam beberapa tahun terakhir, lanskap *Natural Language Processing* (NLP) didominasi oleh kemunculan arsitektur Transformer, khususnya model seperti BERT (Bidirectional Encoder Representations from Transformers) dan variannya. Model-model ini telah menjadi standar baru (de facto standard) dalam berbagai tugas NLP, termasuk analisis sentimen, karena kemampuannya yang superior dalam memahami konteks kalimat secara mendalam, bukan hanya berdasarkan kata per kata [4]. Kemampuan ini memungkinkan model Transformer mencapai akurasi *state-of-the-art* yang sulit ditandingi oleh model-model klasik.

Meskipun demikian, keunggulan model Transformer datang dengan tantangan signifikan, yaitu kompleksitas komputasi yang tinggi, kebutuhan sumber daya (GPU) yang besar, dan waktu inferensi yang lebih lambat. Selain itu, sifatnya yang *black box* membuatnya lebih sulit untuk diinterpretasikan dibandingkan model klasik. Di sisi lain, tantangan utama dalam membangun model klasifikasi yang andal adalah ketersediaan *dataset* berlabel berkualitas tinggi dalam jumlah besar, yang proses pembuatannya secara manual sangat mahal dan memakan waktu.

Penelitian ini memposisikan diri secara unik di tengah lanskap tersebut. Alih-alih hanya membandingkan model *Transformer* dengan model klasik secara langsung, penelitian ini mengusulkan sebuah pendekatan hibrida yang cerdas. Kami memanfaatkan kekuatan pemahaman kontekstual dari model *Transformer* bukan sebagai model klasifikasi akhir, melainkan sebagai 'annotator' atau 'guru' untuk menghasilkan *dataset* ulasan produk *sunscreen* yang berlabel sentimen secara otomatis dan berkualitas tinggi.

Dengan demikian, penelitian ini menjembatani kesenjangan antara kebutuhan akan data latih berkualitas dan penerapan model yang efisien. Kami mengatasi masalah kelangkaan data berlabel (data scarcity) menggunakan teknologi *state-of-the-art* (Transformer), kemudian membangun dan mengevaluasi model-model klasik (LR, NB, SVM) yang lebih ringan, cepat, dan *interpretable* di atas *dataset* tersebut. Pendekatan ini menawarkan solusi pragmatis yang menggabungkan akurasi pelabelan dari model tercanggih dengan efisiensi implementasi dari model yang lebih tradisional.

# BAB III METODOLOGI

## Tahapan Penelitian - Metodologi Data Mining (CRISP-DM)

Metodologi yang digunakan dalam penelitian ini mengacu pada pendekatan **CRISP-DM (Cross Industry Standard Process for Data Mining)**, yang terdiri dari enam tahapan utama. Berikut penjabaran tahapannya:

1. **Akuisisi Data**

Data dikumpulkan dari dua sumber utama, yaitu platform media sosial *Twitter* dan website ulasan produk *Female Daily*. *Dataset* ini terdiri dari total 892 kalimat atau dokumen ulasan terkait produk *Amaterasun Sunscreen Serum*. Format yang digunakan adalah CSV.

1. ***Preprocessing***

Pada tahap ini, data teks dibersihkan dan dipersiapkan dengan langkah-langkah berikut:

1. ***Case* *Folding*:** Mengubah semua huruf menjadi huruf kecil.
2. ***Cleansing*:** Menghapus tanda baca, angka, dan simbol.
3. ***Tokenizing*:** Memecah teks menjadi kata-kata.
4. ***Stopword* *Removal*:** Menghapus kata umum yang tidak memiliki makna penting (seperti “yang”, “dan”, “di”).
5. ***Stemming*:** Mengubah kata ke bentuk dasar.
6. ***Filtering*:** Seleksi teks berdasarkan kriteria panjang atau domain tertentu.

Contoh hasil *preprocessing*:

Sebelum: "Nyaman menggunakan Amaterasun sunscreen bikin putih gak mau yang lain lagi https://t.co/0tdZhulMAE"

Sesudah: "nyaman guna amaterasun sunscreen bikin putih"

1. **Feature Engineering**

Dalam representasi teks menjadi fitur numerik, digunakan dua metode utama:

1. **TF-IDF (Term Frequency - Inverse Document Frequency)**
2. ***Bag of Words* (BoW)**

Ukuran vektor hasil ekstraksi fitur mencapai 2006 dimensi, kemudian dilakukan *feature selection* untuk mereduksi ke 100 fitur yang paling relevan.

1. **Pemodelan**

Tiga algoritma klasifikasi digunakan dalam penelitian ini:

1. ***Logistic Regression***
2. ***Naive Bayes***
3. ***Support Vector Machine* (SVM)**

Masing-masing model diuji menggunakan dua skema validasi:

1. ***Train-Test Split* (80:20)**
2. ***K-Fold Cross Validation***
3. **Evaluasi**

Model dievaluasi menggunakan beberapa metrik evaluasi, yaitu:

1. **Accuracy**
2. ***Precision***
3. ***Recall***
4. ***F1-Score***
5. **Confusion Matrix**

Hasil terbaik diperoleh dari model **SVM** dengan akurasi mencapai 0.8545, *precision* 0.8239, *recall* 0.8545 dan *F1-Score* 0.8084.

## Deskripsi *Dataset*

1. **Sumber Data:** *Twitter* dan website *Female Daily*
2. **Jumlah Data Awal:** 892 dokumen (702 dari Twitter, 190 dari FemaleDaily)
3. **Jumlah Data Setelah *Preprocessing*:** 548 dokumen
4. **Format:** CSV
5. **Tipe Label Sentimen:** Multilabel (Positif, Negatif, Netral)
6. **Contoh Data:**
7. “ini ss terbaik sih buat kulitku yg oily parah...” → Label: **Positif**
8. “lubang tube terlalu besar...” → Label: **Negatif**
9. “Amaterasun sunscreen bikin putih dan gak bikin kulit kering.” → Label: **Netral**

## Algoritma dan Tools yang Digunakan

1. **Algoritma Klasifikasi:**
2. *Logistic Regression*
3. *Naive Bayes*
4. *Support Vector Machine* (SVM)
5. **Tools & Library:**
6. *Python* (Google Colab)
7. *Scikit*-*learn* (sklearn)
8. NLTK / Sastrawi (untuk *preprocessing* teks Bahasa Indonesia)
9. *Matplotlib* & *Seaborn* (untuk visualisasi)
10. *WordCloud* *Library* (untuk peta kata)\

## Evaluasi Kinerja Model

Tabel 3. Tabel Evaluasi Kinerja Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | ***Precision*** | ***Recall*** | ***F1-Score*** |
| ***Logistic Regression*** | 0.8545 | 0.8668 | 0.8545 | 0.7956 |
| ***Naive Bayes*** | 0.8545 | 0.8668 | 0.8545 | 0.7956 |
| **SVM** | 0.8545 | 0.8239 | 0.8545 | 0.8084 |

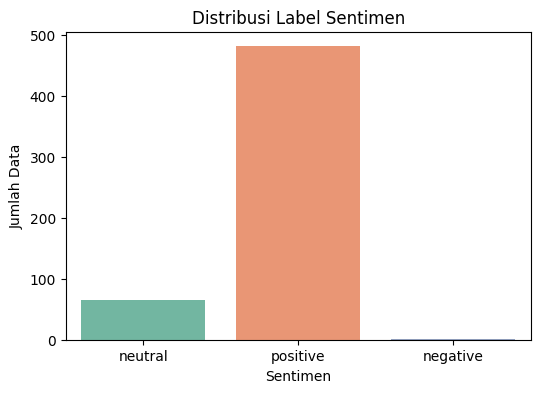
Evaluasi juga disertai dengan visualisasi **confusion matrix**, **WordCloud** untuk sentimen positif dan negatif, **frekuensi kata**, dan distribusi label sentimen.

# BAB IV HASIL DAN PEMBAHASAN

## Visualisasi dan EDA

Eksplorasi data dilakukan untuk memahami karakteristik awal *dataset* ulasan produk *Amaterasun Sunscreen Serum* yang diperoleh dari platform X (Twitter) dan website *Female Daily*.

### Distribusi Label Sentimen



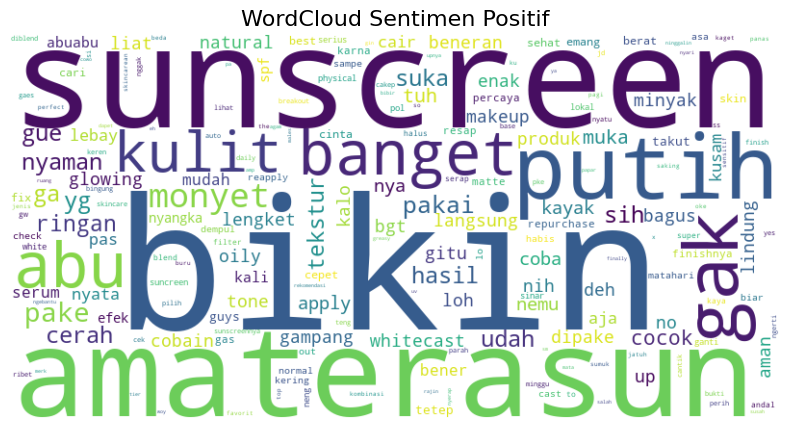
Gambar 4. Distribusi Label Sentimen

Dari total 892 data awal, setelah dilakukan *preprocessing*, hanya 548 data yang layak digunakan. Distribusi sentimen setelah pelabelan menggunakan model Transformer (seperti BERT) adalah sebagai berikut:

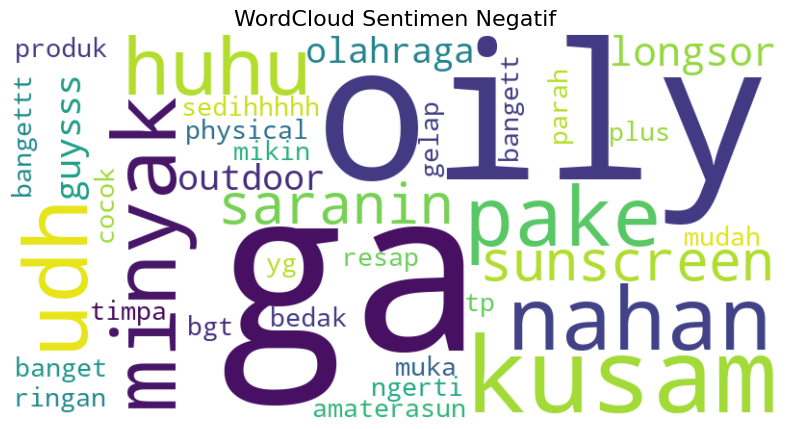
1. Positif: 480 (≈87.6%)
2. Netral: 65 (≈11.9%)
3. Negatif: 3 (≈0.5%)

### WordCloud

WordCloud dibuat untuk masing-masing label sentimen:

1. Positif: Kata-kata dominan seperti “nyaman”, “putih”, “ringan”, dan “cocok” mendominasi.

Gambar 4. WordCloud Positif

1. Negatif: Kata-kata seperti “lengket”, “bau”, “berminyak”, dan “jerawatan” muncul paling sering.

Gambar 4. WordCloud negatif

Visualisasi ini membantu memahami kata-kata umum yang membentuk persepsi pengguna.

### Statistik Deskriptif

Analisis statistik dasar dilakukan pada panjang ulasan dan frekuensi kata:

1. Rata-rata panjang ulasan: 13 kata
2. Maksimal panjang: 42 kata
3. Minimal: 2 kata

## Hasil *Preprocessing* dan Feature Engineering

*Preprocessing* dilakukan menggunakan tahapan *case folding, cleansing, tokenizing, stopword removal, stemming,* dan *filtering.*

Contoh hasil *preprocessing*:

1. Sebelum: "*Nyaman menggunakan Amaterasun sunscreen bikin putih gak mau yang lain lagi...*"
2. Sesudah: "*nyaman guna amaterasun sunscreen bikin putih*"

Selanjutnya, fitur teks direpresentasikan dalam bentuk numerik menggunakan dua pendekatan:

1. TF-IDF: Digunakan sebagai representasi utama karena mempertimbangkan pentingnya kata

dalam dokumen dan keseluruhan korpus.

1. *Bag of Words* (BoW): Digunakan untuk perbandingan, tetapi performa lebih rendah.

Seleksi fitur dilakukan untuk mereduksi dimensi dari 2006 menjadi 100 fitur yang paling relevan, guna menghindari *overfitting* dan mempercepat proses pelatihan.

## Hasil *Preprocessing* dan Pemodelan

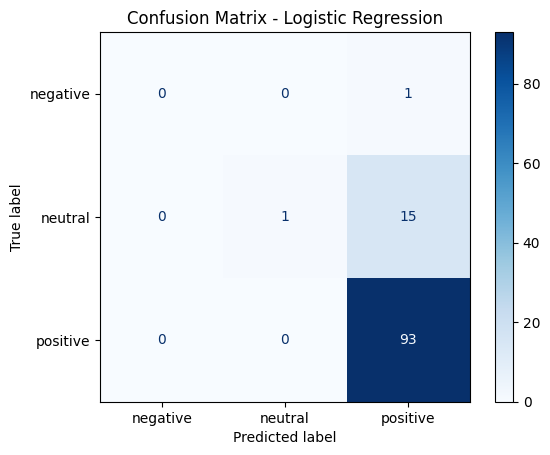
Ketiga algoritma (*Logistic Regression*, *Naive Bayes*, dan SVM) diuji menggunakan skema train‑test split 80:20 serta *k‑fold cross‑validation* (k=5) untuk memastikan kestabilan hasil.

Tabel 4. Hasil Evaluasi Performa Model Klasifikasi Sentimen

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | ***Precision*** | ***Recall*** | ***F1-Score*** |
| ***Logistic Regression*** | 0.8545 | 0.8668 | 0.8545 | 0.7956 |
| ***Naive Bayes*** | 0.8545 | 0.8668 | 0.8545 | 0.7956 |
| **SVM** | 0.8545 | 0.8239 | 0.8545 | 0.8084 |

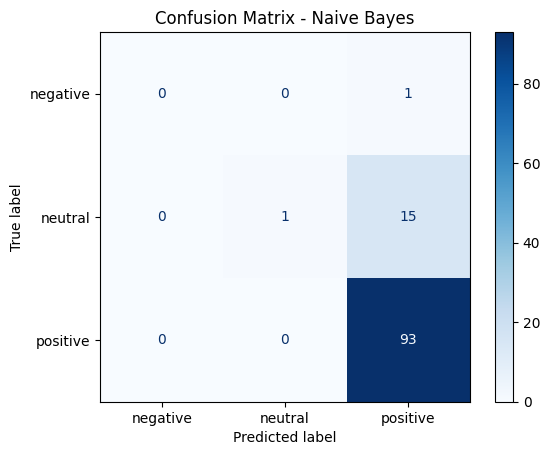
Dari Tabel 4.1 terlihat bahwa ketiga model memiliki nilai akurasi yang sama, yaitu 0.8545. Namun, SVM mencatat nilai *F1-Score* tertinggi (0.8084) dibandingkan dengan *Logistic Regression* dan *Naive Bayes* yang sama-sama memiliki *F1-Score* 0.7956. Hal ini menunjukkan bahwa meskipun ketiga model setara dalam akurasi, SVM memberikan keseimbangan prediksi yang lebih baik antara *precision* dan *recall*.

### Confussion Matrix



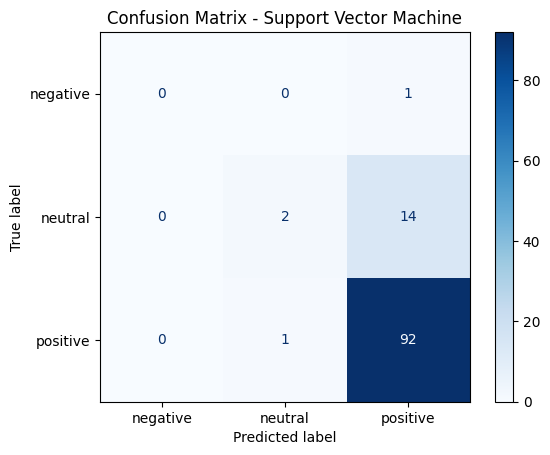
Gambar 4. Confussion Matrix - Logistic Regression

Model *Logistic Regression* menunjukkan performa yang cukup baik dalam mengenali kelas positif. Dari total 93 data positif, seluruhnya berhasil diklasifikasikan dengan benar tanpa ada kesalahan prediksi ke kelas lain. Untuk kelas netral, terdapat 16 data, namun hanya 1 data yang diklasifikasikan dengan benar sebagai netral, sedangkan 15 data lainnya justru diprediksi sebagai positif. Sementara itu, untuk kelas negatif (1 data), model ini salah mengklasifikasikannya sebagai positif. Hal ini mengindikasikan bahwa model *Logistic Regression* cukup andal dalam mengenali kelas positif, tetapi memiliki keterbatasan dalam membedakan kelas netral dan negatif, yang seluruhnya terserap ke kelas mayoritas (positif).



Gambar 4. Confussion Matrix - Naive Bayes

*Naive Bayes* memiliki pola klasifikasi yang sangat mirip dengan *Logistic Regression*. Kelas positif diklasifikasikan dengan sempurna (93 data benar), tanpa salah prediksi. Namun, untuk kelas netral, dari 16 data, hanya 1 data yang terklasifikasi dengan benar, dan 15 data lainnya salah diprediksi sebagai positif. Sama seperti model sebelumnya, kelas negatif (1 data) juga salah diklasifikasikan sebagai positif. Ini menunjukkan bahwa model *Naive Bayes* mengalami bias yang sangat tinggi terhadap kelas positif, dan tidak cukup sensitif terhadap kelas minoritas, terutama jika distribusinya sangat timpang.



Gambar 4. Confussion Matrix – SVM

Model SVM menunjukkan performa klasifikasi yang lebih seimbang dibanding dua model lainnya. Dari 93 data positif, 92 data berhasil diklasifikasikan dengan benar, dan hanya 1 data yang salah diprediksi sebagai netral. Pada kelas netral (16 data), SVM mengklasifikasikan 2 data dengan benar, dan 14 data lainnya masih salah diprediksi sebagai positif. Untuk kelas negatif, yang hanya terdiri dari 1 data, SVM juga tidak berhasil mengklasifikasikannya dengan benar (diprediksi sebagai positif). Meskipun tidak ada model yang mampu mengenali kelas negatif dengan tepat, SVM memiliki kemampuan terbaik dalam mengenali kelas netral, dengan jumlah prediksi benar tertinggi untuk kelas ini, serta kesalahan prediksi paling rendah untuk kelas positif.

## Interpretasi Hasil dan Analisis Komparatif

Ketiga algoritma yang diuji (*Logistic Regression*, *Naive Bayes*, dan SVM) memiliki akurasi yang identik, yaitu sebesar 0.8545. Namun, SVM memberikan hasil terbaik dari segi *F1-Score* (0.8084), yang menunjukkan performa prediksi yang relatif lebih baik dibandingkan model lainnya. Meskipun demikian, semua model menunjukkan kecenderungan untuk mendominasi prediksi pada kelas positif, sehingga distribusi prediksi masih belum sepenuhnya seimbang antar kelas. Meskipun *Naive Bayes* dan *Logistic Regression* memiliki nilai *precision* dan *recall* yang sedikit lebih tinggi dari SVM, *F1-Score* yang lebih baik dari SVM menandakan performa yang lebih konsisten. Oleh karena itu, SVM tetap dipilih sebagai model terbaik dalam eksperimen ini.

## Analisis Keunggulan dan Keterbatasan

Pendekatan yang digunakan dalam penelitian ini memiliki sejumlah keunggulan yang signifikan. Kombinasi antara model modern seperti Transformer untuk pelabelan awal dan model klasik seperti *Logistic Regression*, *Naive Bayes*, dan *Support Vector Machine* (SVM) dalam tahap klasifikasi, menghasilkan sistem yang efisien dan mudah diimplementasikan. Dari ketiga algoritma tersebut, SVM menunjukkan performa terbaik berdasarkan nilai *F1-Score* tertinggi (0.8084). Meskipun nilai metrik tersebut tinggi, performa semua model tetap menunjukkan bias terhadap kelas positif, dengan kemampuan klasifikasi yang terbatas terhadap kelas netral dan negatif. Representasi fitur menggunakan TF-IDF juga terbukti efektif dalam menangani teks berbahasa Indonesia secara ringkas dan informatif. Selain itu, sumber data yang berasal dari dua platform berbeda yaitu *Twitter* dan *Female Daily*, menjadikan *dataset* yang digunakan cukup representatif karena mencerminkan opini pengguna secara lebih beragam.

Namun demikian, terdapat beberapa keterbatasan yang perlu diperhatikan. Ukuran *dataset* bersih yang digunakan berjumlah 548 data, dan terdapat ketimpangan signifikan dalam distribusi label, terutama pada kelas negatif yang hanya berjumlah 3 data. Ketidakseimbangan ini menghambat kemampuan model dalam mengenali kelas minoritas dan berpotensi menurunkan akurasi prediksi pada kelas tersebut. Selain itu, proses pelabelan otomatis menggunakan model *Transformer* juga memiliki potensi bias, sehingga validasi manual tetap diperlukan untuk menjamin kualitas label yang digunakan dalam pelatihan model. Penelitian ini juga belum menerapkan teknik penyeimbangan data seperti SMOTE, yang pada dasarnya dapat membantu meningkatkan kinerja model dalam memprediksi kelas dengan frekuensi rendah dan mengurangi *skew* pada hasil prediksi.

# BAB V KESIMPULAN DAN SARAN

## Kesimpulan

Penelitian ini bertujuan untuk melakukan analisis sentimen terhadap ulasan produk *Amaterasun Sunscreen Serum* dengan pendekatan *machine learning*, menggunakan data yang diperoleh dari platform *Twitter* dan situs *Female Daily*. Proses pelabelan awal dilakukan secara otomatis menggunakan model *Transformer* (seperti BERT), yang memberikan label sentimen tanpa intervensi manual. Data berlabel hasil pelabelan otomatis ini kemudian digunakan sebagai data latih untuk tiga algoritma klasifikasi tradisional: *Logistic Regression*, *Naive Bayes*, dan *Support Vector Machine* (SVM).

Setelah melalui proses *preprocessing* teks dan representasi fitur menggunakan metode TF-IDF, ketiga model menghasilkan nilai akurasi yang sama sebesar 0.8545. Namun, SVM mencatat F1-Score tertinggi sebesar 0.8084, menandakan prediksi yang lebih baik secara umum. Namun demikian, model masih menunjukkan kecenderungan bias terhadap kelas mayoritas (*positif*), dan belum mampu mengenali kelas minoritas seperti *negatif* dengan baik. *Naive Bayes* dan *Logistic Regression* memiliki nilai *F1-Score* yang sedikit lebih rendah (0.7956), dan menunjukkan pola prediksi yang hampir identik. Hasil ini menunjukkan bahwa meskipun ketiga model dapat mengenali pola sentimen secara cukup akurat, SVM tetap menjadi model yang paling konsisten.

Secara umum, kombinasi antara pelabelan otomatis menggunakan *Transformer* dan pemodelan menggunakan algoritma klasik terbukti menjadi pendekatan yang efisien dan dapat diterapkan dalam analisis sentimen ulasan produk lokal berbahasa Indonesia. *Dataset* yang digunakan juga bersifat representatif karena berasal dari dua sumber berbeda dan mencerminkan opini pengguna yang relevan, meskipun masih memiliki keterbatasan pada distribusi kelas yang sangat timpang, terutama jumlah data negatif yang sangat sedikit.

## Jawaban atas Rumusan Masalah

Model *Transformer* (seperti BERT) dimanfaatkan sebagai annotator otomatis yang memberikan label sentimen pada *dataset* tanpa perlu pelabelan manual. Pendekatan ini mempercepat proses pengumpulan data berlabel dari ulasan pengguna yang sebelumnya tidak memiliki anotasi sentimen, serta memungkinkan klasifikasi awal terhadap opini konsumen dengan efisiensi tinggi dan potensi akurasi yang baik.

Ketiga algoritma klasifikasi yaitu *Logistic Regression*, *Naive Bayes*, dan *Support Vector Machine* (SVM) menunjukkan performa yang cukup baik dalam membangun model klasifikasi sentimen, dengan nilai akurasi yang sama sebesar 85.45%. Namun demikian, terdapat perbedaan pada nilai F1-Score masing-masing model, yang mencerminkan kemampuan model dalam menjaga keseimbangan antara precision dan recall, meskipun semua model cenderung mendominasi prediksi pada kelas positif. Di antara ketiganya, SVM memiliki nilai *F1-Score* tertinggi, menandakan performa prediksi yang lebih stabil dan seimbang antar kelas.

Berdasarkan evaluasi menyeluruh terhadap metrik performa seperti akurasi, precision, recall, dan F1-Score, algoritma Support Vector Machine (SVM) dinilai sebagai model paling optimal dalam mengklasifikasikan sentimen ulasan produk Amaterasun. Namun, evaluasi ini tetap perlu dibarengi pemahaman bahwa distribusi prediksi masih belum merata antar kelas. Kemampuannya dalam menangani data berdimensi tinggi dan memberikan prediksi yang relatif seimbang menjadikannya pilihan terbaik di antara ketiga algoritma yang diuji.

## Saran

Untuk pengembangan penelitian selanjutnya, terdapat beberapa aspek yang dapat ditingkatkan guna memperoleh hasil klasifikasi yang lebih optimal. Pertama, menambah jumlah data ulasan dan memperluas sumber data dapat meningkatkan kemampuan generalisasi model, terutama pada kelas minoritas seperti sentimen negatif. Kedua, validasi manual atau semi-otomatis terhadap hasil pelabelan sentimen tetap perlu dilakukan untuk mengurangi kemungkinan bias dari model Transformer. Ketiga, penerapan teknik penyeimbangan data seperti SMOTE atau oversampling dapat membantu memperbaiki distribusi kelas dan meningkatkan akurasi pada kelas dengan jumlah data yang rendah. Selain itu, eksplorasi terhadap algoritma lain seperti Random Forest atau XGBoost, serta pemanfaatan representasi fitur berbasis embedding modern seperti Word2Vec atau BERT embedding, dapat menjadi arah pengembangan yang menjanjikan untuk penelitian lanjutan dalam analisis sentimen teks berbahasa Indonesia.

# DAFTAR PUSTAKA

[1] S. Nabil, J. Elbouhdidi, and M. Chkouri, “Sentiment Analysis Of Amazon’s Reviews Using Machine Learning Algorithms,” 2021.

[2] M. Haroon, Z. Alam, R. Kousar, J. Ahmad, and F. Nasim, “Sentiment Analysis of Customer Reviews on E-commerce Platforms: A Machine Learning Approach,” *Bulletin of Business and Economics (BBE)*, 2024, doi: 10.61506/01.00480.

[3] M. M. Aziz, M. D. Purbalaksono, and A. Adiwijaya, “Method comparison of Naïve Bayes, *Logistic Regression*, and SVM for Analyzing Movie Reviews,” *Building of Informatics, Technology and Science (BITS)*, 2023, doi: 10.47065/bits.v4i4.2644.

[4] M. K. Chenglerayen and Dr. S. R. Raja, “From Reviews to Results: Leveraging Amazon Feedback for Product Evolution,” *Int J Sci Res Sci Eng Technol*, 2024, doi: 10.32628/ijsrset2411458.

[5] P. K. Sari and R. R. Suryono, “Komparasi algoritma *Support Vector Machine* dan Random Forest untuk analisis sentimen metaverse,” *Jurnal Mnemonic*, vol. 7, no. 1, pp. 31–39, 2024.

[6] R. Merdiansah, S. Siska, and A. A. Ridha, “Analisis sentimen pengguna X Indonesia terkait kendaraan listrik menggunakan IndoBERT,” *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI)*, vol. 7, no. 1, pp. 221–228, 2024.

[7] D. Sjoraida, B. Guna, and D. Yudhakusuma, “Analisis Sentimen Film Dirty Vote Menggunakan BERT (Bidirectional Encoder Representations from Transformers),” *Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi)*, vol. 8, pp. 393–404, Jul. 2024, doi: 10.35870/jtik.v8i2.1580.

[8] C. Yang and C. Huang, “Natural language processing (NLP) in aviation safety: Systematic review of research and outlook into the future,” *Aerospace*, vol. 10, no. 7, p. 600, 2023.

[9] F. K. Khaiser, A. Saad, and C. Mason, “Sentiment analysis of students’ feedback on institutional facilities using text-based classification and natural language processing (NLP),” *Journal of Language and Communication*, vol. 10, no. 1, pp. 101–111, 2023.

[10] I. Oraif, “Natural Language Processing (NLP) and EFL learning: A case study based on Deep Learning,” *Journal of Language Teaching and Research*, vol. 15, no. 1, pp. 201–208, 2024.

[11] M. Fadhil, I. M. M. Matin, and S. Nurpadillah, “Analisis Klasifikasi Sentimen terhadap Penurunan Covid-19 Menggunakan Metode Na\"\ive Bayes dan *Support Vector Machine* pada Platform X App,” *Fuse-teknik Elektro*, vol. 3, no. 2, pp. 85–93, 2023.

[12] A. A. Permana, W. A. Noviyanto, and D. A. Kristiyanti, “Sentimen Analisis Opini Masyarakat Terhadap UMKM Pada Media Sosial Twitter Dengan Metode Na\"\ive Bayes Classifier,” *Jurnal Minfo Polgan*, vol. 12, no. 1, pp. 163–170, 2023.

[13] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, “Applications of *Support Vector Machine* (SVM) learning in cancer genomics,” *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, 2018.

# LAMPIRAN

1. Lampiran A – *Dataset* dan Informasi Terkait
2. Lampiran A1 – Deskripsi *Dataset*

Sumber Data: Twitter atau X dan Ulasan pada Website Female Daily

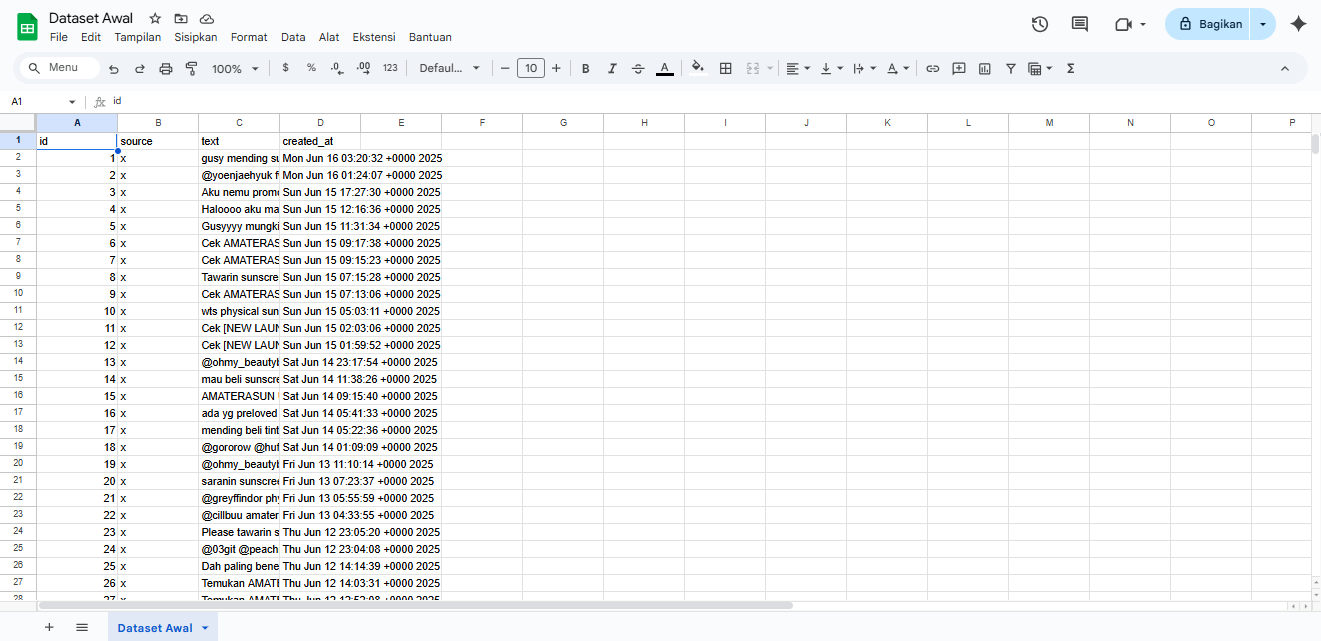
Jumlah Data: 892 Data

Jumlah Atribut: 4

Deskripsi Atribut:

* Atribut 1: id
* Atribut 2: source
* Atribut 3: text
* Atribut 4: created\_at

1. Lampiran A2 – Contoh *Dataset* Mentah (Raw)



Gambar. Gambar Dataset Mentah atau Raw

1. Lampiran B – Proses *Preprocessing*
2. Lampiran B1 – Data Cleaning

Langkah-langkah pembersihan:

Penanganan nilai kosong: -

Duplikasi data: -

Outlier: -

1. Lampiran B2 – Transformasi Data

Jenis transformasi:

Normalisasi/Standarisasi: -

Encoding: -

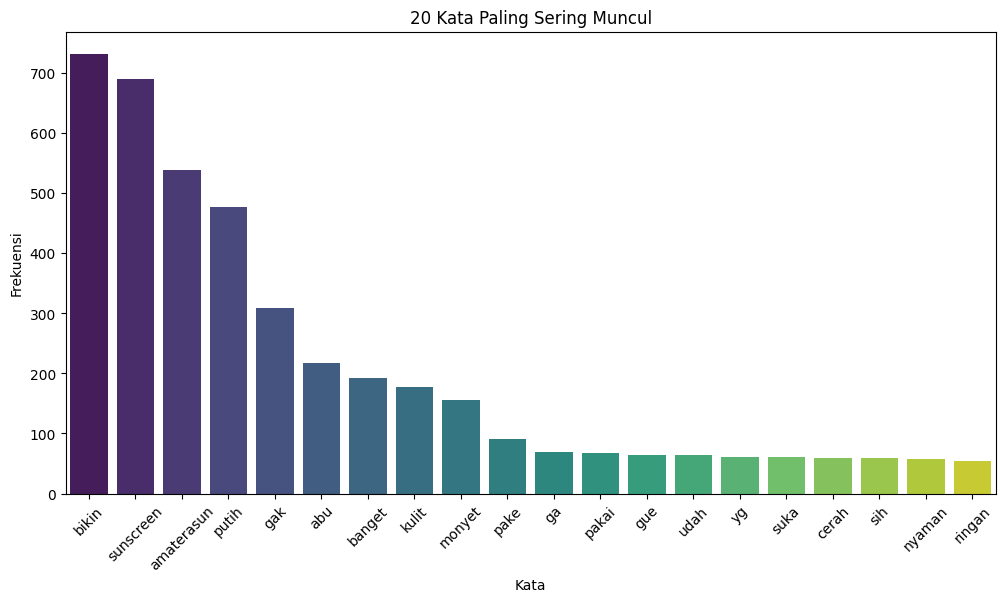
Binning/Discretization: -

1. Lampiran C – Eksplorasi Data & Visualisasi (EDA)
2. Lampiran C1 – Statistik Deskriptif

Tampilkan tabel statistik: mean, median, modus, min, max, std

1. Lampiran C2 – Grafik dan Visualisasi

(Tambahkan visualisasi EDA: histogram, boxplot, scatterplot, heatmap)



1. Lampiran D – Pemodelan dan Evaluasi
2. Lampiran D1 – Rincian Model

Model yang digunakan:

1. Logistic Regression

2. Naïve Bayes

3. Support Vector Machine (SVM)

Parameter model:

1. Logistic Regression:

* penalty 'l2'
* dual False
* tol 0.0001
* C 1.0
* fit\_intercept True
* intercept\_scaling 1
* class\_weight None
* random\_state None
* solver 'lbfgs'
* max\_iter 1000
* multi\_class 'deprecated'
* verbose 0
* warm\_start False
* n\_jobs None
* l1\_ratio None

1. Naïve Bayes (MultinomialNB):

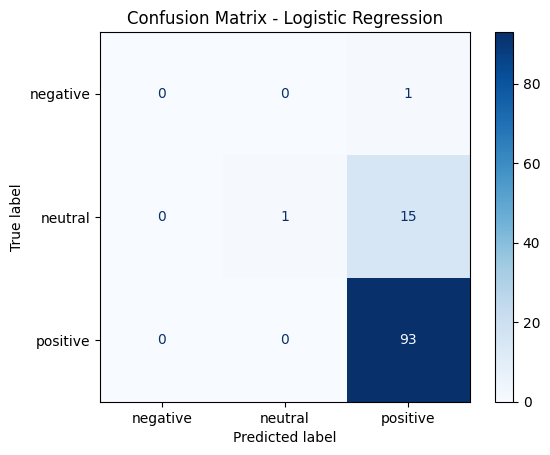
* alpha 1.0
* force\_alpha True
* fit\_prior True
* class\_prior None

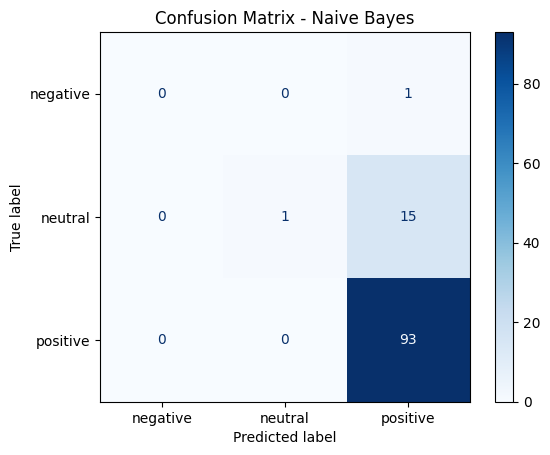
1. Support Vector Machine:

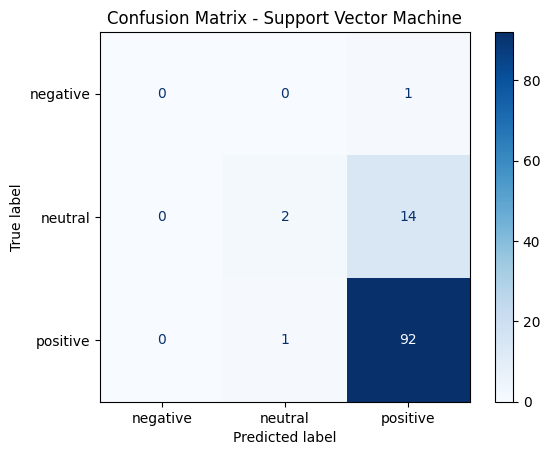
* penalty 'l2'
* loss 'squared\_hinge'
* dual 'auto'
* tol 0.0001
* C 1.0
* multi\_class 'ovr'
* fit\_intercept True
* intercept\_scaling 1
* class\_weight None
* verbose 0
* random\_state None
* max\_iter 1000

1. Lampiran D2 – Hasil Evaluasi Model

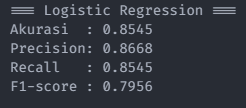
* Confusion Matrix

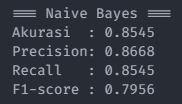


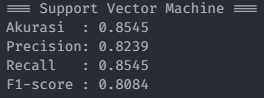




* Classification Report (Accuracy, *Precision*, *Recall*, F1)







1. Lampiran E – Kode Program
2. Lampiran E1 – Script Python/R/Notebook



Kode untuk split data traning sebesar 0.2 dan juga kode untuk menginisialisasikan model Logistik Regresi.

[Link Notebook](https://github.com/kelompok-4-dataminig/tugas-besar-datamining-kelompok4/blob/main/notebook/modelling.ipynb)

1. Lampiran E2 – Struktur Folder Proyek

/data

/notebook

/report

/src