LAPORAN PROYEK

ANALISIS SENTIMEN ULASAN

APLIKASI TIX ID

MENGGUNAKAN METODE KLASIFIKASI

SUPPORT VECTOR MACHINE

Diajukan Untuk Memenuhi Tugas Kuliah

**”Penambangan Data”**

Dosen Pengampu :

Cucun Very Angkoso, S.T., MT



**Disusun Oleh :**

Nama : Harits Putra Junaidi

NIM : 230411100003

Prodi : Teknik Informatika

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK**

**UNIVERSITAS TRUNOJOYO MADURA**

**2024/2025**

# PENGANTAR

## Latar Belakang

Smartphone telah menjadi salah satu kebutuhan utama di era modern yang semakin maju. Seiring dengan itu, perkembangan teknologi internet juga menjadi simbol komunikasi tanpa batas, baik dari segi ruang, jarak, maupun waktu [1]. Oleh karena itu, kombinasi antara smartphone dan internet menjadikan berbagai aktivitas serta pekerjaan manusia jauh lebih efisien dibandingkan sebelumnya.

Untuk menunjang kemudahan dalam menyelesaikan pekerjaan, penggunaan smartphone tidak dapat dipisahkan dari peran aplikasi mobile. Saat ini, tren teknologi mengarah pada dominasi aplikasi mobile, khususnya media sosial yang digunakan oleh sekitar 92 juta orang, atau setara dengan 32 persen dari total populasi di Indonesia [2].

Hadirnya aplikasi-aplikasi mobile seperti Shopee, Gojek, dan Grab sangat membantu masyarakat dalam melakukan kegiatan sehari-hari. Salah satu aplikasi pemesanan tiket bioskop online tersebut yaitu TIX ID. TIX ID merupakan salah satu perusahaan layanan atau aplikasi pemesanan tiket film bioskop secara online yang dibuat oleh PT Nusantara Elang Sejahtera. Aplikasi ini memberikan layanan seperti memesan tiket film dari bioskop favorit, memilih film online untuk dibeli atau sewa, dan membeli voucher untuk menonton film streaming online [3].

Dengan hadirnya aplikasi mobile TIX ID, menjadi gerbang utama konsumen untuk mengakses layanan hiburan. Aplikasi ini pada situs layanan Google Play Store telah diunduh sebanyak 10 juta lebih dengan rating 4,8 dan 515 ribu lebih ulasan. Sebagai platform pemesanan tiket bioskop terdepan di Indonesia, melayani jutaan pengguna dan menerima ribuan feedback melalui ulasan Google Play Store setiap bulannya.

Ulasan pengguna merupakan aset informasi yang sangat berharga karena mengandung pemahaman mendalam tentang kepuasan pelanggan, permasalahan teknis, preferensi fitur, dan ekspektasi pengguna. Namun, volume ulasan yang mencapai ribuan per minggu menciptakan tantangan signifikan dalam hal analisis dan response time.

Keterlambatan dalam mengidentifikasi dan merespons sentimen negatif dapat berdampak pada penurunan rating aplikasi, kehilangan pengguna, dan damage terhadap brand reputation. Di era digital yang kompetitif, kemampuan untuk memberikan response yang cepat dan tepat sasaran terhadap feedback pengguna menjadi competitive advantage yang krusial.

Analisis sentimen menggunakan machine learning, khususnya Support Vector Machine (SVM), menawarkan solusi untuk mengotomatisasi proses klasifikasi ulasan berdasarkan sentimen positif, negatif, atau netral. SVM dipilih karena Beberapa penelitian telah menunjukkan bahwa SVM dapat menghasilkan akurasi yang lebih tinggi dalam klasifikasi ulasan dibandingkan algoritma lainnya, seperti Naive Bayes atau K-Nearest Neighbors [4].

## Pernyataan Masalah

* Bagaimana cara mengklasifikasikan sentimen ulasan pengguna aplikasi TIX ID secara otomatis?
* Implementasi Metode klasifikasi yang paling efektif untuk menganalisis sentimen ulasan aplikasi TIX ID?
* Bagaimana performa Support Vector Machine dalam mengklasifikasikan sen- timen ulasan?

## Tujuan Penelitian

* Bagaimana cara mengklasifikasikan sentimen ulasan pengguna aplikasi TIX ID secara otomatis?
* Implementasi Metode klasifikasi yang paling efektif untuk menganalisis sentimen ulasan aplikasi TIX ID?
* Bagaimana performa Support Vector Machine dalam mengklasifikasikan sen- timen ulasan?

## Manfaat Penelitian

* Membantu tim pengembang TIX ID memahami feedback pengguna secara real-time
* Mempercepat proses analisis ulasan pelanggan
* Memberikan dasar untuk perbaikan aplikasi berdasarkan sentimen pengguna

# BUSINESS UNDERSTANDING

## Latar Belakang Bisnis

TIX ID adalah aplikasi pemesanan tiket bioskop online yang populer di Indonesia. Aplikasi ini memungkinkan pengguna untuk memesan tiket bioskop dari berbagai jaringan cinema secara digital melalui smartphone.

Sebagai platform digital, TIX ID menerima banyak ulasan dari pengguna di Google Play Store setiap harinya. Ulasan ini berisi feedback tentang pengalaman pengguna, mulai dari kepuasan terhadap fitur aplikasi, keluhan teknis, hingga saran perbaikan.

Dalam era digital saat ini, ulasan pengguna memiliki peran penting dalam mempengaruhi reputasi aplikasi dan keputusan download pengguna baru. Oleh karena itu, memahami sentimen dari ulasan pengguna menjadi hal yang krusial bagi pengembangan dan peningkatan kualitas aplikasi.

## Permasalahan Bisnis

* TIX ID menerima ratusan ulasan setiap minggu di Google Play Store. Volume yang besar ini membuat analisis manual menjadi tidak efisien dan memakan waktu yang lama.
* Tanpa sistem otomatis, tim sulit untuk mengidentifikasi dengan cepat ulasan mana yang bersifat positif, negatif, atau netral. Hal ini menyebabkan keterlambatan dalam merespons feedback pengguna.
* Keterlambatan dalam memahami sentimen pengguna dapat berdampak pada:
* Penurunan rating aplikasi di Play Store
* Kehilangan kesempatan untuk memperbaiki masalah dengan cepat
* Menurunnya kepuasan pengguna secara keseluruhan

## Tujuan Bisnis

* Mengembangkan sistem yang dapat secara otomatis mengklasifikasikan sentimen ulasan pengguna menjadi kategori positif, atau negatif.
* Meningkatkan efisiensi dalam proses analisis feedback pengguna sehingga tim dapat lebih cepat dalam memberikan respons dan perbaikan.
* Memberikan pemahaman yang dapat digunakan untuk pengambilan keputusan dalam pengembangan fitur dan perbaikan aplikasi berdasarkan sentiment analysis.

## Kriteria Keberhasilan Bisnis

* Model klasifikasi sentimen mampu mencapai tingkat akurasi minimal 80% dalam mengklasifikasikan ulasan pengguna.
* Sistem dapat memproses dan mengklasifikasikan ulasan dalam waktu yang jauh lebih singkat dibandingkan analisis manual.
* Berhasil mengimplementasikan model Support Vector Machine untuk analisis sentimen ulasan aplikasi TIX ID dengan hasil yang dapat diandalkan untuk tim pengembang.

## Manfaat Operasional

Sistem yang dikembangkan dapat memberikan manfaat nyata dalam:

* Mempercepat identifikasi feedback negatif
* Membantu prioritas perbaikan berdasarkan sentimen pengguna
* Mendukung pengambilan keputusan untuk pengembangan aplikasi

# PEMBAHASAN

## Understanding

### Sumber Data

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle dengan link: <https://www.kaggle.com/datasets/ahmadseloabadi/tix-id-app-reviews-from-google-play-store>

Dataset berupa file CSV dengan nama scrapped\_TIX ID\_EN.csv yang berisi ulasan aplikasi TIX ID dari Google Play Store.

### Deskripsi Dataset

Dataset memiliki 11 kolom dengan rincian sebagai berikut:

|  |  |  |
| --- | --- | --- |
| **Kolom** | **Tipe Data** | **Deskripsi** |
| reviewId | String | ID unik untuk setiap ulasan |
| userName | String | Nama pengguna yang memberikan ulasan |
| userImage | String | URL gambar profil pengguna |
| content | String | Isi ulasan dalam bahasa Indonesia |
| score | Integer | Rating yang diberikan (1-5) |
| thumbsUpCount | Integer | Jumlah like pada ulasan |
| reviewCreatedVers | String | Versi aplikasi saat ulasan dibuat |
| at | Datetime | Waktu ulasan dibuat |
| replyContent | String | Balasan dari developer (jika ada) |
| repliedAt | Datetime | Waktu balasan dibuat |
| appVersion | String | Versi aplikasi |

### Eksplorasi Data Awal

Berdasarkan analisis dalam kode yang diberikan:

* Dataset memiliki total 28.247 data ulasan (namun data yang akan dipakai hanya 10.000 data)
* Kolom utama yang digunakan untuk analisis adalah content dan score
* Distribusi rating menunjukkan variasi sentimen dari pengguna
* Terdapat ulasan dalam bahasa Indonesia yang memerlukan preprocessing khusus

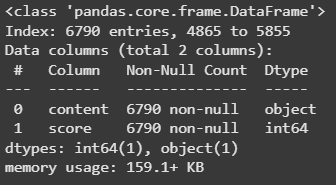
## Data Preparation

### Pembersihan Data

Tahapan preprocessing yang dilakukan berdasarkan implementasi dalam kode:

* **Penanganan Data Duplicat**

Menghapus baris dengan nilai duplicat pada kolom content



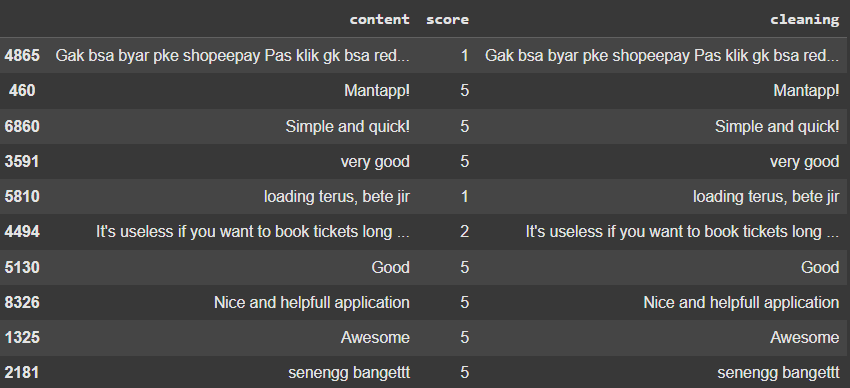
Gambar 3. 1 Hasil Penanganan Data Duplikat

Yang awalnya punya 10.000 data, tapi setelah menghapus duplikat pada kolom content, tinggal 6.790 data unik. Dimana, content dan score masing-masing punya 6790 data (tidak ada nilai kosong).

* **Text Preprocessing**

**Cleaning**

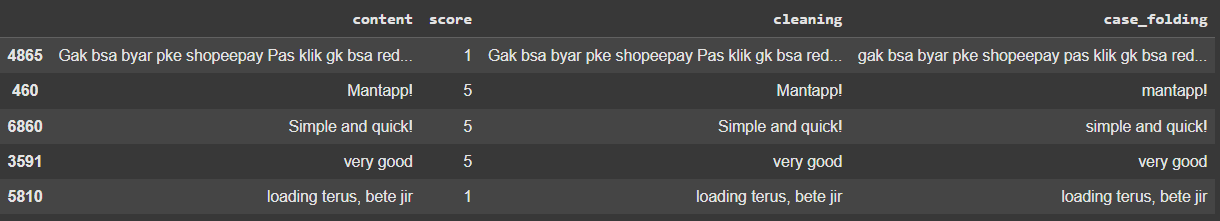
Dalam tahap pembersihan data (*data cleaning*), dilakukan proses penghapusan elemen-elemen yang tidak dibutuhkan agar teks menjadi lebih bersih dan siap untuk dianalisis. Proses ini mencakup penghapusan karakter URL, tag HTML, emoji, angka, serta simbol-simbol khusus yang tidak memberikan makna terhadap analisis teks [5]. Pembersihan ini penting untuk mengurangi noise dalam data dan memastikan bahwa hasil analisis lebih akurat serta representatif terhadap konteks sebenarnya.



Gambar 3. 2 Hasil Data Cleaning

**Case Folding**

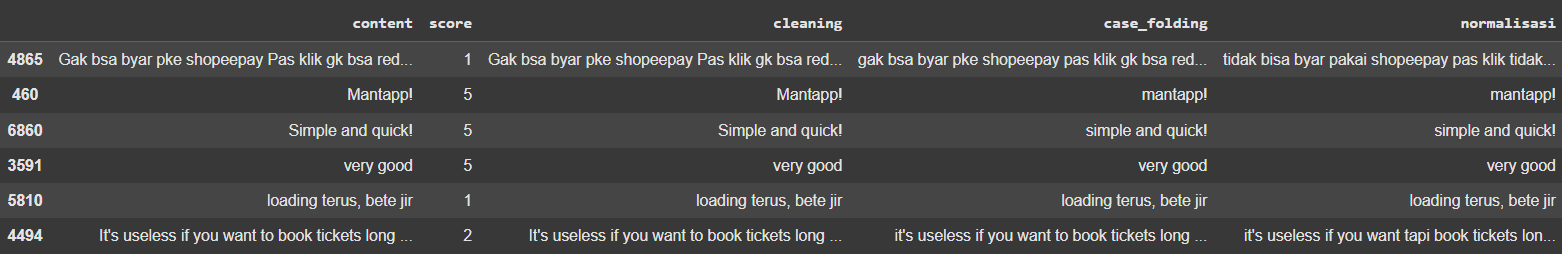
Case folding adalah proses mengubah seluruh huruf dalam teks menjadi huruf kecil untuk menyeragamkan data [6]. Langkah ini mencegah perbedaan kata akibat kapitalisasi, sehingga analisis teks menjadi lebih konsisten dan akurat.



Gambar 3. 3 Hasil Case Folding

**Normalisasi**

Normalisasi merupakan proses mengganti semua kata tidak baku atau tidak sesuai ejaan kata baku yang sesuai dengan kaidah bahasa Indonesia. Tahapan ini penting untuk meningkatkan konsistensi data teks, sehingga kata-kata yang memiliki arti sama namun bentuk tidak bakunya berbeda dapat dikenali sebagai satu entitas yang seragam.



Gambar 3. 4 Hasil Normalisasi

**Tokenization**

Tokenization adalah proses memecah teks menjadi bagian-bagian kecil yang disebut token, seperti kata atau frasa. Tahapan ini bertujuan untuk memudahkan analisis teks dengan mengidentifikasi setiap elemen penyusunnya secara terpisah.



Gambar 3. 5 Gambar Hasil Tekenization

**Stopwords Removal**

Stopwords removal adalah proses menghapus kata-kata umum yang tidak memiliki makna penting dalam analisis, seperti "ke", "yang", atau "biasanya" [7]. Dalam proses ini, digunakan daftar stopwords bahasa Indonesia untuk menyaring kata yang tidak memberikan makna penting terhadap pemahaman konteks teks.



Gambar 3. 6 Hasil Stopword Removal

**Stemming**

Stemming adalah Mengurangi variasi kata yang sama dengan menghapus imbuhan [8]. Dalam tahap ini, digunakan **Sastrawi stemmer** untuk bahasa Indonesia agar kata turunan seperti "makanan", "memakan", atau "dimakan" dapat dikembalikan ke bentuk dasar "makan", sehingga analisis teks menjadi lebih efektif dan terfokus pada makna inti.



Gambar 3. 7 Hasil Stemming

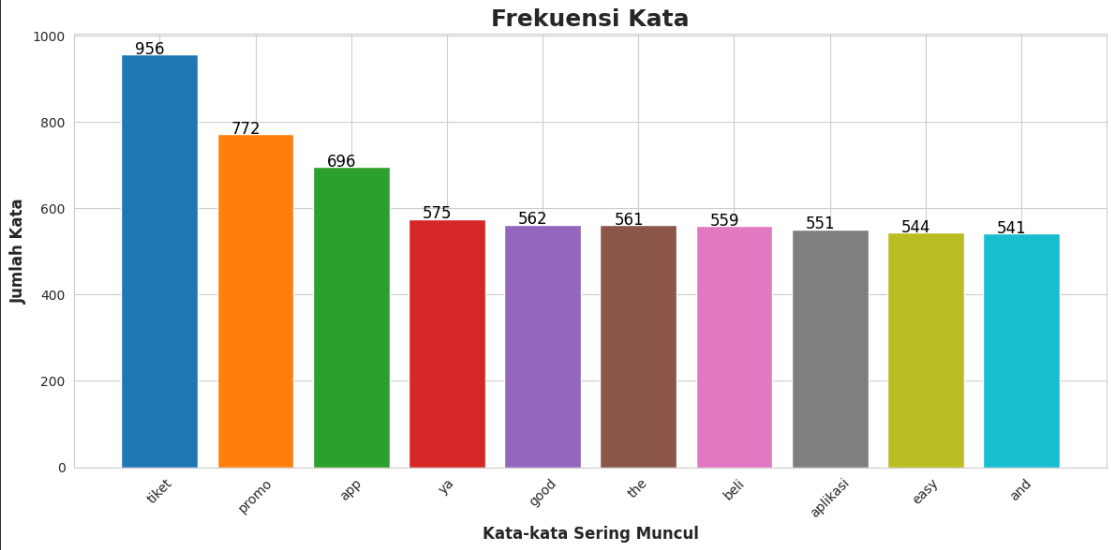
* **Hasil Preprocessing**

Proses preprocessing berhasil dilakukan dengan hasil Wordcloud dan Diagram kata yang sering muncul sebagai berikut:



Gambar 3. 8 WordCloud Setelah Preprocessing

Gambar 3. 9 WordCloud Sebelum Preprocessing



Gambar 3. 10 Frekuensi Kata

### Lebeling Sentimen

Pelabelan sentimen menggunakan pendekatan berbasis leksikon dengan memanfaatkan kamus kata sentimen bahasa Indonesia dari repository InSet (<https://github.com/fajri91/InSet>). Metode ini dipilih untuk memberikan hasil yang objektif dan konsisten.

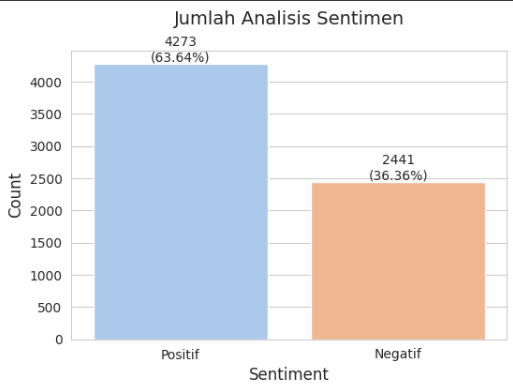
* 1. **Algoritma Penentuan Sentimen**

Sentimen ditentukan melalui perhitungan frekuensi kata:

* Menghitung jumlah kata positif dan negatif dalam setiap teks
* Jika kata positif > kata negatif → **Positif**
* Jika kata positif < kata negatif → **Negatif**
  1. **Sumber Leksikon**
* **positive.tsv**: Kamus kata-kata berkonotasi positif bahasa Indonesia
  + - * **negative.tsv**: Kamus kata-kata berkonotasi negatif bahasa Indonesia
  1. **Hasil Distribusi**

Proses pelabelan menghasilkan dataset dengan distribusi:

* **Sentimen Positif**: 63.64%
* **Sentimen Negatif**: 36.36%



Gambar 3. 11 Jumlah Analisi Sentimen

### Feature Extraction

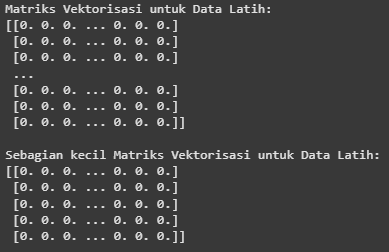
* TF-IDF Vectorization

TF-IDF (Term Frequency-Inverse Document Frequency) adalah metode untuk mengkonversi teks menjadi representasi numerik yang dapat diproses oleh algoritma machine learning.

**Proses Implementasi**

1. **Fitting**: Vectorizer mempelajari vocabulary dari data training dan menghitung nilai IDF
2. **Transform Training**: Mengkonversi data training menjadi matriks TF-IDF
3. **Transform Testing**: Mengaplikasikan vocabulary yang sama pada data testing

**Hasil Vektorisasi**

****

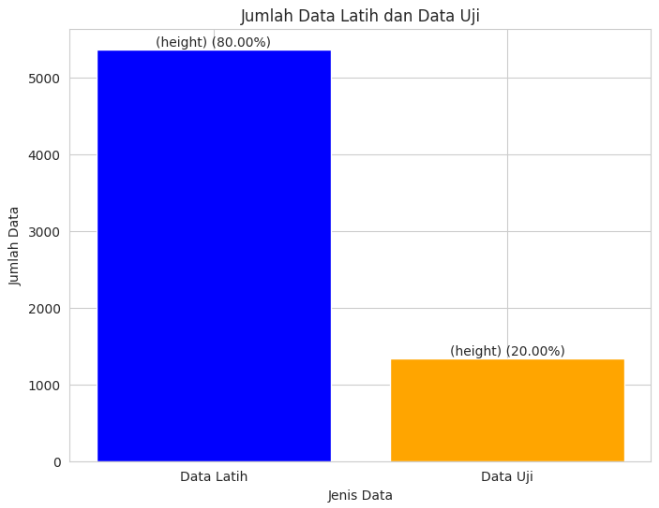
Gambar 3. 12 Hasil Vektorisasi

* **Format Matrix:** Matriks berukuran [n\_samples, n\_features]
* **Sparse Nature:** Sebagian besar nilai adalah 0 karena setiap dokumen hanya mengandung subset kecil dari total vocabulary
* **Representasi:** Setiap baris merepresentasikan satu ulasan, setiap kolom merepresentasikan satu kata dalam vocabulary

### Data Splitting

Total jumlah data setelah dilakukan Preprocessing, Menghapus data yang nilainya null, dan Labelling Data adalah sebanyak 6.714. Data tersebut akan dibagi menjadi:

* Training set: 80% (5.371 Data)
* Testing set: 20% (1.343 Data)



Gambar 3. 13 Diagram Jumlah Data Latih dan data Uji

## Modelling

### Algoritma yang Dipilih

* **Support Vector Machine (SVM)**

SVM dipilih sebagai algoritma utama dengan pertimbangan:

* + - * Efektif untuk klasifikasi teks
      * Dapat menangani high-dimensional data
      * Robust terhadap overfitting
      * Performa baik pada dataset dengan feature yang banyak

### Konfigurasi Model

Penelitian ini menggunakan algoritma Support Vector Machine (SVM) dari library scikit-learn untuk melakukan klasifikasi sentimen pada ulasan aplikasi TIX ID.Model SVM dikonfigurasi dengan parameter:

* **Kernel**: Linear - dipilih untuk menangani data teks yang umumnya linearly separable dalam high-dimensional space
* **C (Regularization)**: 1.0 (default) - parameter regularisasi untuk mengontrol trade-off antara margin maksimum dan error klasifikasi
* **Gamma**: 'scale' (default) - tidak relevan untuk linear kernel
* **Class Weight**: None - menggunakan distribusi natural dari data

### Pipeline Training

1. Preprocessing teks ulasan

* Data Cleaning, Case Folding, Normalisasi Kata, Tokenizing, Stopword Removal, dan Steaming

1. Data Splitting

* Pembagian data menjadi training 80% dan testing set 20%

1. Feature Engineering

* TF-IDF Vectorization untuk mengkonversi teks menjadi numerical features

1. Training model SVM

* Fitting model SVM dengan kernel linear pada data training
* Learning pattern dari TF-IDF features untuk memprediksi sentiment labels

1. Model Evaluation

* Mengukur performance dengan metrics seperti accuracy, precision, recall

## Evaluasi

### Matrik Evaluasi

Model dievaluasi menggunakan beberapa metrik:

* **Confusion Matrix**

Menampilkan prediksi benar dan salah untuk setiap kelas sentimen.

* **TP (True Positives)** = 312 → Model benar memprediksi **positif** saat datanya memang positif.
* **FN (False Negatives)** = 191 → Model salah memprediksi **negatif**, padahal datanya positif.
* **FP (False Positives)** = 84 → Model salah memprediksi **positif**, padahal datanya negatif.
* **TN (True Negatives)** = 756 → Model benar memprediksi **negatif** saat datanya memang negatif.
  + - * **Classification Report**

**Sentimen Negatif:**

* **Precision (0.79)**: 79% prediksi negatif adalah benar
* **Recall (0.62)**: Model berhasil mendeteksi 62% dari seluruh sentimen negatif
* **F1-Score (0.69)**: Performa gabungan cukup baik
* **Support (503)**: 503 ulasan negatif dalam dataset testing

**Sentimen Positif:**

* **Precision (0.80)**: 80% prediksi positif adalah benar
* **Recall (0.90)**: Model berhasil mendeteksi 90% dari seluruh sentimen positif
* **F1-Score (0.85)**: Performa gabungan sangat baik
* **Support (840)**: 840 ulasan positif dalam dataset testing

**Performa Overall:**

* **Accuracy (0.80)**: 80% prediksi secara keseluruhan benar
* **Macro Average (0.77)**: Rata-rata sederhana dari kedua kelas
* **Weighted Average (0.79)**: Rata-rata berbobot sesuai jumlah instance
* **Accuracy Score**

Persentase prediksi yang benar dari total prediksi adalah 0.80 (80%)

### Hasil Evaluasi

Berdasarkan implementasi dalam kode, hasil evaluasi menunjukkan:

Tabel Performa Model SVM

|  |  |  |  |
| --- | --- | --- | --- |
| **Kelas** | **Precision** | **Recall** | **F1-Score** |
| Negatif | 0.79 | 0.62 | 0.69 |
| Positif | 0.80 | 0.90 | 0.85 |
| **Weighted Avg** | 0.79 | 0.80 | 0.79 |

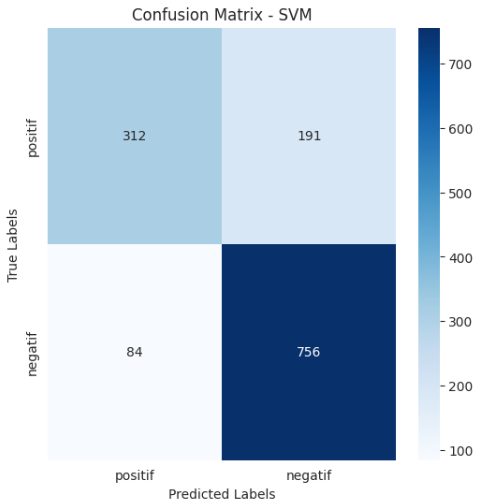
### Analisis Performa

* Model menunjukkan performa yang baik dalam mengklasifikasikan sentimen
* Kelas dengan performa terbaik: **Sentimen Positif**
* Tantangan : **Sentimen Negatif** menghadapi tantangan dalam
  + **Low Recall (0.62)**: 38% sentimen negatif tidak terdeteksi (False Negatives tinggi)
  + Imbalanced Performance: Gap signifikan antara recall positif (0.90) vs negatif (0.62)
  + Business Impact: Risiko missed opportunity dalam mendeteksi keluhan pelanggan

### Visualisasi Hasil

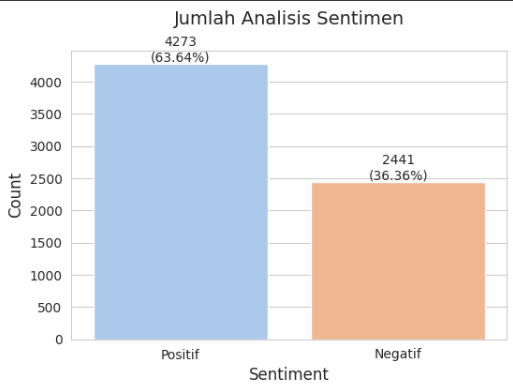
Visualisasi yang dihasilkan meliputi:

* **Confusion matrix** 
  + **TP (True Positives)** = 312 → Model benar memprediksi **positif** saat datanya memang positif.
  + **FN (False Negatives)** = 191 → Model salah memprediksi **negatif**, padahal datanya positif.
  + **FP (False Positives)** = 84 → Model salah memprediksi **positif**, padahal datanya negatif.
  + **TN (True Negatives)** = 756 → Model benar memprediksi **negatif** saat datanya memang negatif.

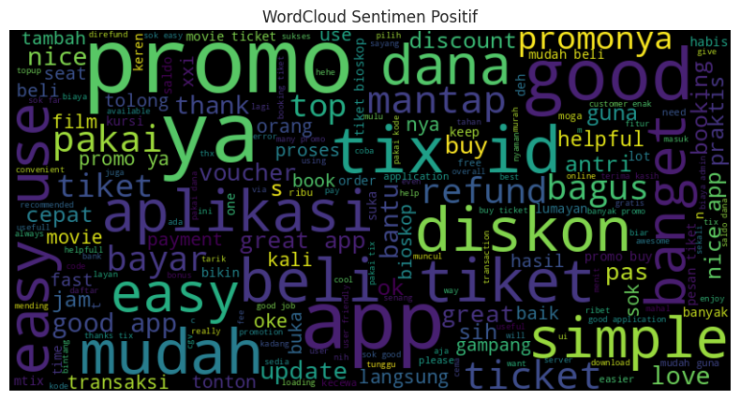


Gambar 3. 14 Counfusion Matix – SVM

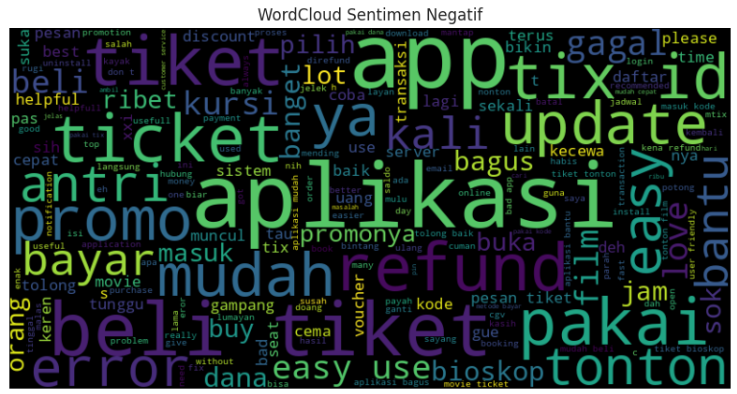
* **Distribusi sentimen dalam dataset**
  + **Sentimen Positif** : Count: 4,273 ulasan (63.64%)
  + **Sentimen Negatif :** Count: 2,441 ulasan (36.36%)



Gambar 3. 15 Distribusi sentimen dalam dataset

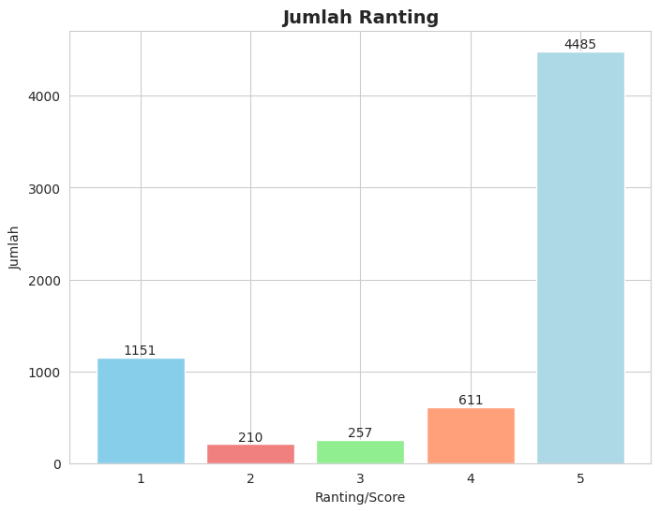
* **Wordcloud untuk setiap kelas sentimen**
  + Wordcloud Sentimen Positif

Gambar 3. 16 Wordcloud Sentimen Positif

* + Wordcloud Sentimen Negatif

Gambar 3. 17 Wordcloud Sentimen Negatif

* **Jumlah Rating Aplikasi TIX ID**

****

Gambar 3. 18 Jumlah Rating Aplikasi TIX ID

Berdasarkan visualisasi distribusi rating aplikasi TIX ID, terlihat pola polarisasi ekstrem dimana rating 5 bintang mendominasi dengan 4,485 ulasan (66.8%), diikuti rating 1 bintang sebanyak 1,151 ulasan (17.1%), sementara rating tengah relatif sedikit. Hal ini menunjukkan mayoritas pengguna sangat puas (75.9% memberikan rating 4-5), namun terdapat segmen pengguna yang sangat tidak puas (20.2% memberikan rating 1-2). Distribusi ini konsisten dengan analisis sentimen 63.64% positif, mengkonfirmasi TIX ID memiliki performa baik secara keseluruhan dengan ruang perbaikan untuk mengatasi keluhan pengguna yang mengalami masalah serius.

# KESIMPULAN

# DAFTAR PUSTAKA

[1] M. C. Ngulum, & A. D. Indriyanti,“Evaluasi Kualitas Website Simontasi Unesa Menggunakan Metode Webqual Dan Importance Performance Analysis (Ipa),”Journal of Informatics and Computer Science (JINACS), vol.2, no. 1, hal. 38-43, 2020.

[2] Yolanovia, D., & Indriyanti, A. D. (2021). Evaluasi User Experience Aplikasi TIX ID Menggunakan Metode Heuristic Evaluation. *Journal of Emerging Information System and Business Intelligence (JEISBI)*, *2*(3), 8-13.

[3] Nadhif, M. R., Brata, D. W., & Rahayudi, B. (2022). Analisis Sentimen Data Ulasan Pengguna Aplikasi TIX ID di Indonesia pada Google Play Store menggunakan Support Vector Machine. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *6*(8), 3932-3937.

[4] N. Khoirunnisaa, K. Nabila, N. Kesuma, S. Setiawan, A. Yunizar, and P. Yusuf, “KLASIFIKASI TEKS ULASAN APLIKASI NETFLIX PADA GOOGLE PLAY STORE MENGGUNAKAN ALGORITMA NAÏVE BAYES DAN SVM,” SKANIKA Sist. Komput. dan Tek. Inform., vol. 7, no. 1, pp. 64–73, Jan. 2024, doi: 10.36080/SKANIKA.V7I1.3138.

[5] Alhaq, Z., Mustopa, A., & Santoso, J. D. (2021). Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter.

[6] Dwi Cahyo, M. P., Widodo, & Prasetya Adhi, B. (2019). Kinerja Algoritma Support Vector Machine dalam MenentukanKebenaran Informasi Banjir di Twitter. PINTER : Jurnal Pendidikan Teknik Informatika Dan Komputer, 3(2), 116–121.

[7] Fitriyana, V., Hakim, L., Candra Rini Novitasari, D., Hanif Asyhar, A., Studi Matematika, P., Sains Dan Teknologi, F., Sunan Ampel Surabaya, U., & Timur, J. (2023). Analisis Sentimen Ulasan Aplikasi Jamsostek Mobile Menggunakan Metode Support Vector Machine. In Jurnal Buana Informatika (Vol. 14, Issue 1).

[8] N. Octaviani Faomasi Daeli, “Sentiment Analysis on Movie Reviews Using Information Gain and KNearest Neighbor,” J. Data Sci. Its Appl., 2020.