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

1 Business Understanding

1.1 Latar Belakang

Industri entertainment digital di Indonesia mengalami pertumbuhan pesat, dengan aplikasi mobile menjadi gerbang utama konsumen untuk mengakses layanan hiburan. TIX ID, 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 insight 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.

1.2 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 sentimen ulasan?

1.3 Tujuan Penelitian

- Mengembangkan model klasifikasi sentimen otomatis untuk ulasan aplikasi TIX ID
- Mengimplementasikan algoritma Support Vector Machine untuk analisis sentimen
- Mengevaluasi performa model yang dikembangkan

1.4 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

2 Understanding

2.1 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 data

Dataset berupa file CSV dengan nama scrapped_TIX ID_EN.csv yang berisi ulasan aplikasi TIX ID dari Google Play Store.

2.2 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

2.3 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 khu- sus

3 Data Preparation

3.1 Pembersihan Data

Tahapan preprocessing yang dilakukan berdasarkan implementasi dalam kode:

3.1.1 Penanganan Data Duplicat

- Menghapus baris dengan nilai duplicat pada kolom content

```
<class 'pandas.core.frame.DataFrame'>
Index: 6790 entries, 4865 to 5855
Data columns (total 2 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   content  6790 non-null   object
 1   score    6790 non-null   int64
dtypes: int64(1), object(1)
memory usage: 159.1+ KB
```

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).

3.1.2 Text Preprocessing

1. **Cleaning**: Menghapus karakter url, html, emoji, angka, dan simbol

	content	score	cleaning
4865	Gak bsa byar pke shopeepay Pas klik gk bsa red...	1	Gak bsa byar pke shopeepay Pas klik gk bsa red...
460	Mantapp!	5	Mantapp!
6860	Simple and quick!	5	Simple and quick!
3591	very good	5	very good
5810	loading terus, bete jir	1	loading terus, bete jir
4494	It's useless if you want to book tickets long ...	2	It's useless if you want to book tickets long ...
5130	Good	5	Good
8326	Nice and helpfull application	5	Nice and helpfull application
1325	Awesome	5	Awesome
2181	senengg bangettt	5	senengg bangettt

2. **Case Folding**: Mengubah semua teks menjadi huruf kecil

	content	score	cleaning	case_folding
4865	Gak bsa byar pke shopeepay Pas klik gk bsa red...	1	Gak bsa byar pke shopeepay Pas klik gk bsa red...	gak bsa byar pke shopeepay pas klik gk bsa red...
460	Mantapp!	5	Mantapp!	mantapp!
6860	Simple and quick!	5	Simple and quick!	simple and quick!
3591	very good	5	very good	very good
5810	loading terus, bete jir	1	loading terus, bete jir	loading terus, bete jir

3. **Normalisasi :** Mengganti semua kata tidak baku menjadi kata baku

	content	score	cleaning	case_folding	normalisasi
4865	Gak bsa byar pke shopeepay Pas klik gk bsa red...	1	Gak bsa byar pke shopeepay Pas klik gk bsa red...	gak bsa byar pke shopeepay pas klik gk bsa red...	tidak bisa byar pakai shopeepay pas klik tidak...
460	Mantapp!	5	Mantapp!	mantapp!	mantapp!
6860	Simple and quick!	5	Simple and quick!	simple and quick!	simple and quick!
3591	very good	5	very good	very good	very good
5810	loading terus, bete jir	1	loading terus, bete jir	loading terus, bete jir	loading terus, bete jir
4494	It's useless if you want to book tickets long ...	2	It's useless if you want to book tickets long ...	it's useless if you want to book tickets long ...	it's useless if you want tapi book tickets lon...

4. **Tokenization**: Memecah teks menjadi token individual

	content	score	cleaning	case_folding	normalisasi	tokenize
4865	Gak bsa byar pke shopeepay Pas klik gk bsa red...	1	Gak bsa byar pke shopeepay Pas klik gk bsa red...	gak bsa byar pke shopeepay pas klik gk bsa red...	tidak bisa byar pakai shopeepay pas klik tidak...	[tidak, bisa, byar, pakai, shopeepay, pas, kli...
460	Mantapp!	5	Mantapp!	mantapp!	mantapp!	[mantapp!]
6860	Simple and quick!	5	Simple and quick!	simple and quick!	simple and quick!	[simple, and, quick!]
3591	very good	5	very good	very good	very good	[very, good]
5810	loading terus, bete jir	1	loading terus, bete jir	loading terus, bete jir	loading terus, bete jir	[loading, terus,, bete, jir]

5. **Stopwords Removal**: Menghapus kata-kata yang tidak bermakna menggu- nakan stopwords bahasa Indonesia

	content	score	cleaning	case_folding	normalisasi	tokenize	stopword removal
4865	Gak bsa byar pke shopeepay Pas klik gk bsa red...	1	Gak bsa byar pke shopeepay Pas klik gk bsa red...	gak bsa byar pke shopeepay pas klik gk bsa red...	tidak bisa byar pakai shopeepay pas klik tidak...	[tidak, bisa, byar, pakai, shopeepay, pas, kli...	[byar, pakai, shopeepay, pas, klik, redirect....
460	Mantapp!	5	Mantapp!	mantapp!	mantapp!	[mantapp!]	[mantapp!]
6860	Simple and quick!	5	Simple and quick!	simple and quick!	simple and quick!	[simple, and, quick!]	[simple, and, quick!]
3591	very good	5	very good	very good	very good	[very, good]	[very, good]
5810	loading terus, bete jir	1	loading terus, bete jir	loading terus, bete jir	loading terus, bete jir	[loading, terus,, bete, jir]	[loading, terus,, bete, jir]

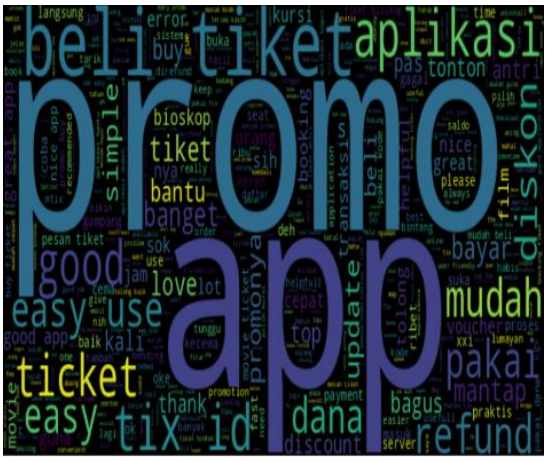
6. **Stemming**: Mengubah kata ke bentuk dasarnya menggunakan Sastrawi stem- mer

	content	score	cleaning	case_folding	normalisasi	tokenize	stopword removal	steming_data
4865	Gak bsa byar pke shopeepay Pas klik gk bsa red...	1	Gak bsa byar pke shopeepay Pas klik gk bsa red...	gak bsa byar pke shopeepay pas klik gk bsa red...	tidak bisa byar pakai shopeepay pas klik tidak...	[tidak, bisa, byar, pakai, shopeepay, pas, kli...	[byar, pakai, shopeepay, pas, klik, redirect....	byar pakai shopeepay pas klik redirect gk canc...
460	Mantapp!	5	Mantapp!	mantapp!	mantapp!	[mantapp!]	[mantapp!]	mantapp
6860	Simple and quick!	5	Simple and quick!	simple and quick!	simple and quick!	[simple, and, quick!]	[simple, and, quick!]	simple and quick
3591	very good	5	very good	very good	very good	[very, good]	[very, good]	very good
5810	loading terus, bete jir	1	loading terus, bete jir	loading terus, bete jir	loading terus, bete jir	[loading, terus,, bete, jir]	[loading, terus,, bete, jir]	loading terus bete jir

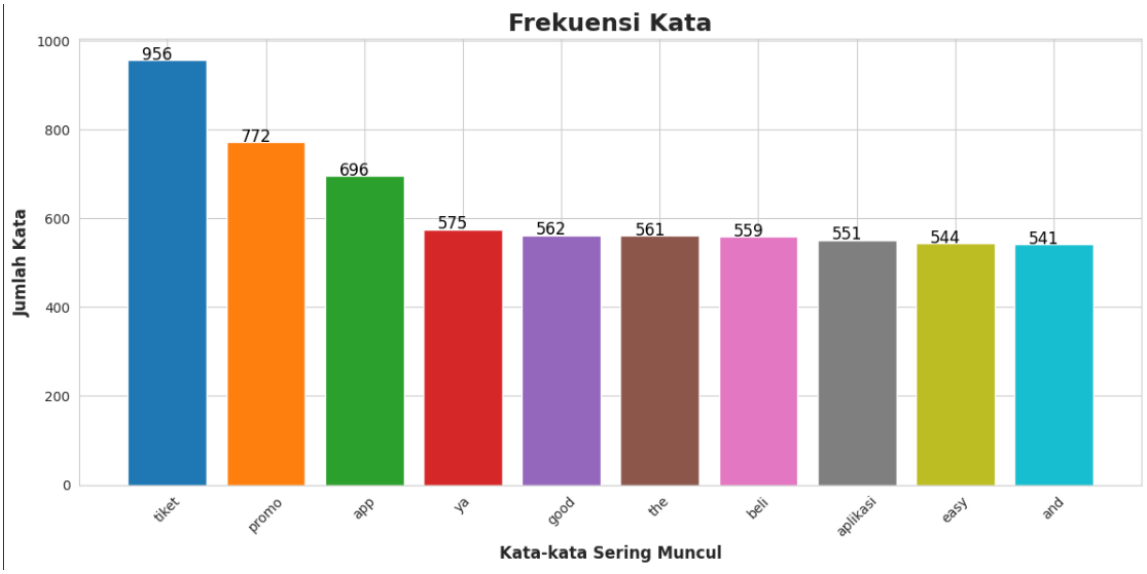
3.1.3 Hasil Preprocessing

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

Gambar 1: WordCloud Sebelum Preprocessing



Gambar 2: WordCloud Setelah Preprocessing



Gambar 3: Diagram Frekuensi Kata (Kata-kata Sering Muncul) setelah Preprocessing

3.2 Labeling 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.

3.2.1 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**
- Untuk kasus ambiguitas (jumlah sama atau tidak ada kata sentimen), dilakukan pengisian bergantian untuk menjaga keseimbangan dataset

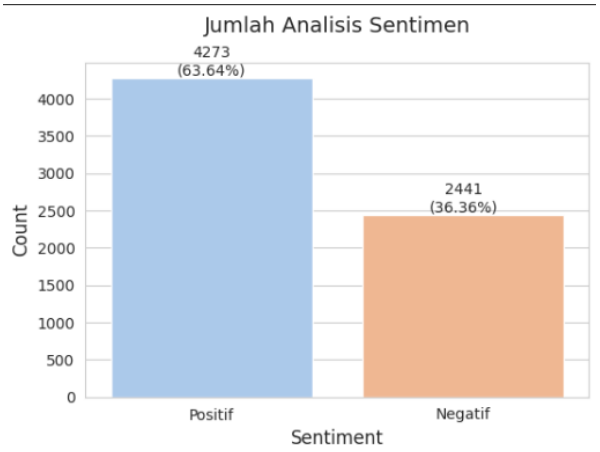
3.2.2 Sumber Leksikon

- **positive.tsv**: Kamus kata-kata berkonotasi positif bahasa Indonesia
- **negative.tsv**: Kamus kata-kata berkonotasi negatif bahasa Indonesia

3.2.3 Hasil Distribusi

Proses pelabelan menghasilkan dataset dengan distribusi:

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



Gambar 4: Diagram Analisis Sentimen

3.3 Feature Extraction

3.3.1 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

```
Matriks Vektorisasi untuk Data Latih:
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]

Sebagian kecil Matriks Vektorisasi untuk Data Latih:
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

Gambar 5 Hasil Vektorisasi

- **Format Matrix:** Matriks berukuran [n_samples, n_features]
- **Sparse Nature:** Sebagian besar nilai adalah 0 karena setiap dokumen

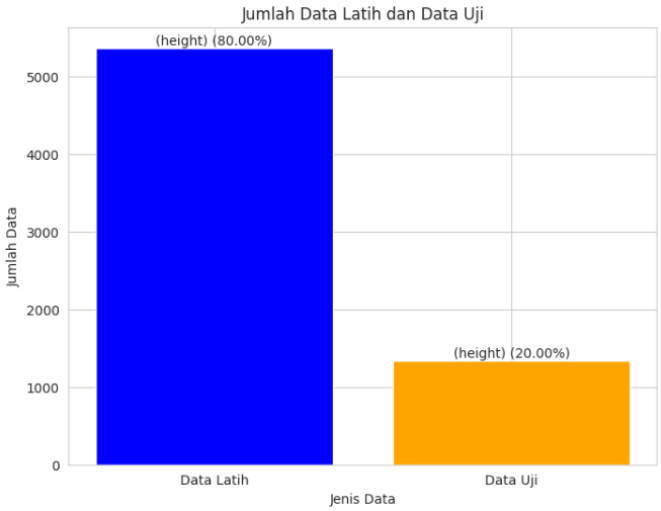
hanya mengandung subset kecil dari total vocabulary

- **Representasi:** Setiap baris merepresentasikan satu ulasan, setiap kolom merepresentasikan satu kata dalam vocabulary

3.4 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 6: Diagram Jumlah Data Latih dan data Uji

4 Modelling

4.1 Algoritma yang Dipilih

4.1.1 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

4.2 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

4.3 Pipeline Training

1. Preprocessing teks ulasan
    - Data Cleaning, Case Folding, Normalisasi Kata, Tokenizing, Stopword Removal, dan Steaming
2. Data Splitting
    - Pembagian data menjadi training 80% dan testing set 20%
3. Feature Engineering
    - TF-IDF Vectorization untuk mengkonversi teks menjadi numerical features
4. Training model SVM
    - Fitting model SVM dengan kernel linear pada data training
    - Learning pattern dari TF-IDF features untuk memprediksi sentiment labels
5. Model Evaluation
    - Mengukur performance dengan metrics seperti accuracy, precision, recall

5 Evaluasi

5.1 Metrik Evaluasi

Model dievaluasi menggunakan beberapa metrik:

5.1.1 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.

5.1.2 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

5.1.3 Accuracy Score

Persentase prediksi yang benar dari total prediksi adalah 0.80 atau 80%

5.2 Hasil Evaluasi

Berdasarkan implementasi dalam kode, hasil evaluasi menunjukkan:

Tabel 2: 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

5.3 Analisis Performa

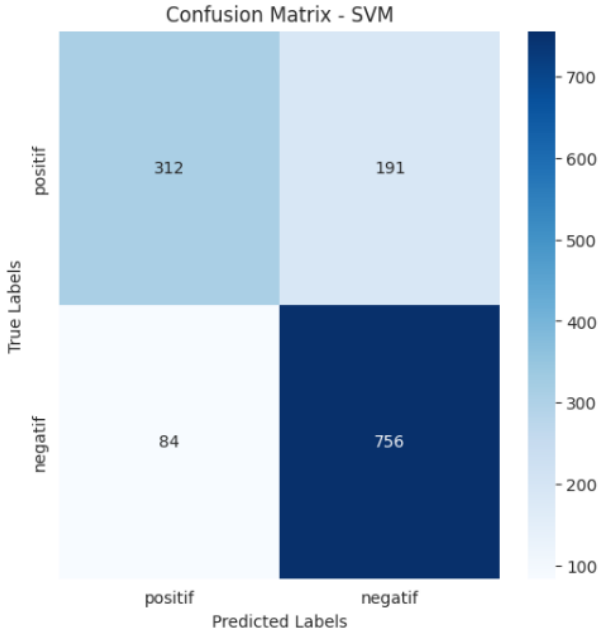
- Model menunjukkan performa yang baik dalam mengklasifikasikan sentimen

- 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

5.4 Visualisasi Hasil

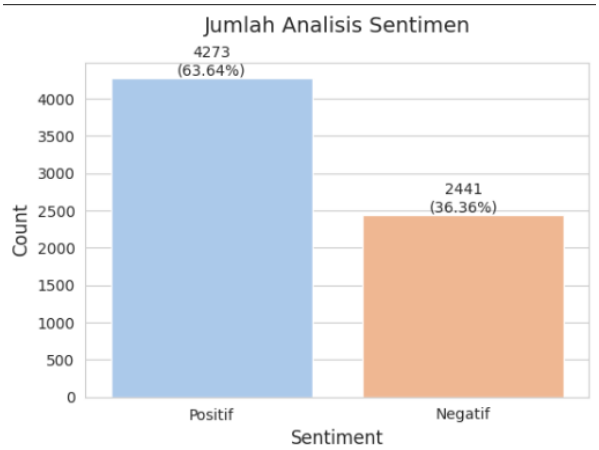
Visualisasi yang dihasilkan meliputi:

- 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.



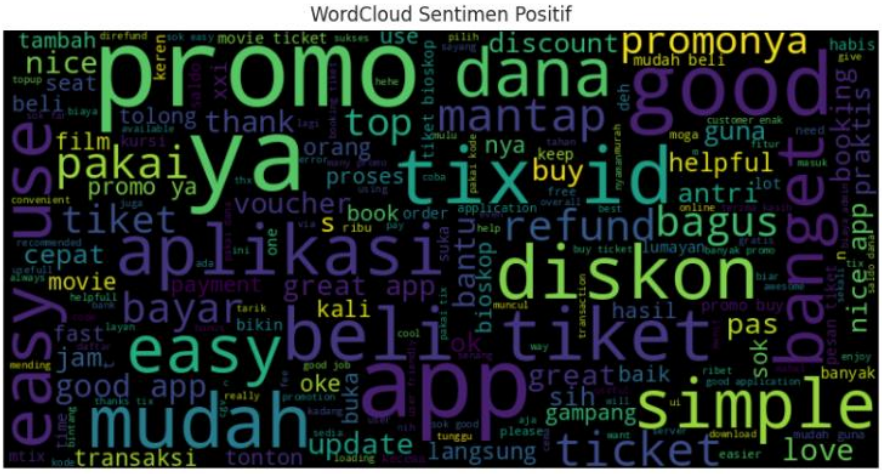
Gambar 7 Counfusion Matix – SVM

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



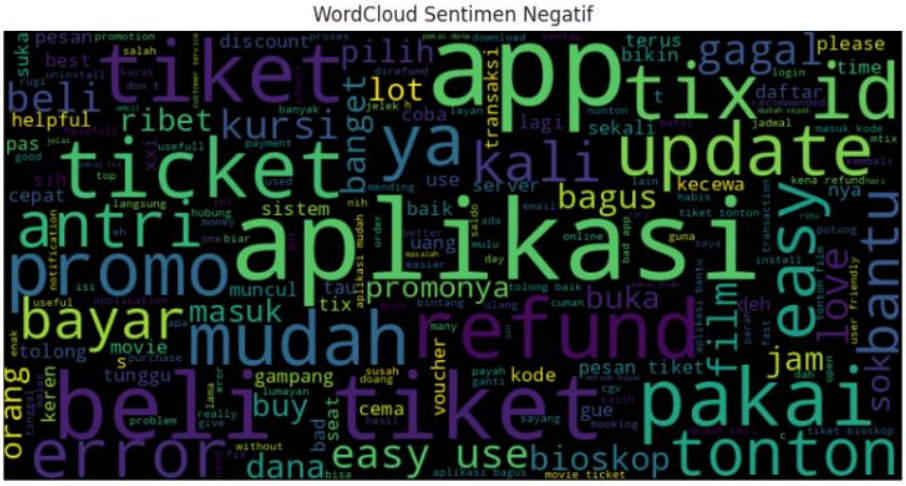
Gambar 8: Distribusi sentimen dalam dataset

- Wordcloud untuk setiap kelas sentimen
  - Wordcloud Sentimen Positif



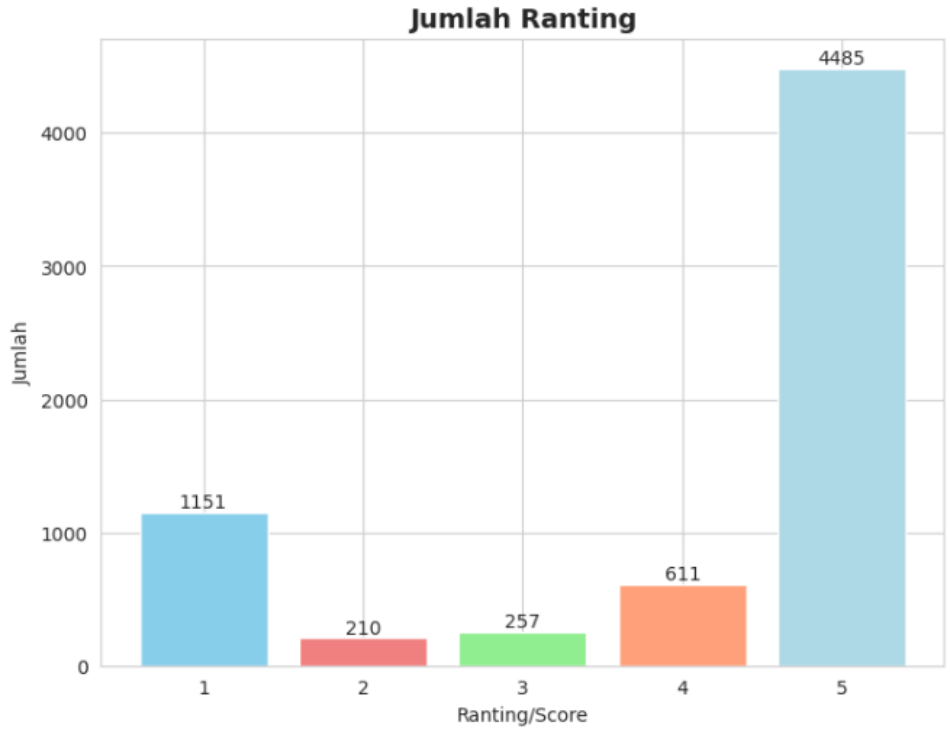
Gambar 9: Wordcloud Sentimen Positif

  - Wordcloud Sentimen Negatif



Gambar 10: Wordcloud Sentimen Negatif

- **Jumlah Rating Aplikasi TIX ID**



Gambar 11: 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.

6 Deployment

6.1 Target End-User

Model yang dikembangkan ditujukan untuk:

6.1.1 Tim Product Management TIX ID

- Monitoring sentimen pengguna secara real-time
- Identifikasi tren feedback pengguna
- Pengambilan keputusan berbasis data sentimen

6.1.2 Tim Customer Service

- Prioritisasi penanganan keluhan berdasarkan sentimen negatif
- Respons cepat terhadap ulasan negatif
- Monitoring kepuasan pelanggan

6.1.3 Tim Development

- Identifikasi bug atau fitur yang bermasalah
- Prioritas pengembangan fitur baru
- Evaluasi dampak update aplikasi terhadap sentimen pengguna

6.2 Pentingnya Proyek

Proyek ini Penting karena dapat :

- Merespons lebih cepat terhadap feedback pengguna
- Otomatisasi analisis ribuan ulasan pengguna
- Pengurangan waktu analisis manual dari hari menjadi menit
- Peningkatan retensi pengguna melalui perbaikan yang tepat sasaran
- Optimalisasi strategi marketing berdasarkan sentimen positif