

## Analisis Sentimen Data Ulasan Pengguna Aplikasi TIX ID di Indonesia pada Google Play Store menggunakan *Support Vector Machine*

Muhammad Razan Nadhif<sup>1</sup>, Dwija Wisnu Brata<sup>2</sup>, Bayu Rahayudi<sup>3</sup>

Program Studi Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Brawijaya  
Email: <sup>1</sup>razannadhif@student.ub.ac.id, <sup>2</sup>wisnubrata@ub.ac.id, <sup>3</sup>ubay1@ub.ac.id

### Abstrak

*Review* pada Google Play Store merupakan salah satu fitur yang digunakan untuk memberikan suatu penilaian terhadap suatu aplikasi. 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 disewa, dan membeli *voucher* untuk menonton film *streaming online*. Namun rating yang disertai dengan berbagai ulasan bukan berarti bahwa keinginan atau masalah pengguna terpenuhi dan berhenti untuk melakukan evaluasi dalam peningkatan pelayanan kepada pengguna. Untuk itu diperlukan analisis sentimen yang dapat mengklasifikasikan review sebagai sentimen pengguna. Dalam penelitian ini dilakukan tahap *scraping* untuk pengumpulan data ulasan pengguna aplikasi, dilanjutkan tahap *text preprocessing* untuk mengolah data dengan menyeleksi data dan mengubahnya menjadi data yang lebih terstruktur. Data hasil *text preprocessing* dilakukan pembobotan kata dengan menggunakan metode Term Frequency – Inverse Document Frequency (TF-IDF). Kemudian dilakukan klasifikasi sentimen menggunakan algoritma Support Vector Machine (SVM). Hasil terbaik yang didapatkan dengan algoritma SVM untuk pengujian sentimen 2 kelas dengan data tidakseimbang dengan data positif 274 dan data negatif 100 menggunakan nilai data latih dan data uji 90%:10%, total data dengan 374 data, menggunakan parameter nilai  $C = 10$ , percobaan *cross validation*  $K=10$  dan penggunaan kernel *linear*. Hasil yang didapatkan untuk nilai rata - rata *accuracy* 91%, *precision* 94%, *recall* 83% dan *f-measure* 86%. Hasil terbaik yang didapatkan dengan algoritma SVM untuk pengujian sentimen 2 kelas dengan data seimbang dengan data positif 150 dan data negatif 150 menggunakan nilai data latih dan data uji 90%:10%, total data dengan 300 data, menggunakan parameter nilai  $C = 0.5$ , percobaan *cross validation*  $K=10$  dan penggunaan kernel *linear*. Hasil pengujian pada data seimbang mendapatkan hasil akurasi tertinggi yakni 94%, *precision* 94%, *recall* 94% dan *f-measure* 94% pada metode *Support Vector Machine* klasifikasi 2 sentimen.

**Kata kunci:** Analisis Sentimen, Scraping, Text Preprocessing, TF-IDF, Support Vector Machine, Google Play Store

### Abstract

*Review on the Google Play Store is one of the features used to provide an assessment of an application. TIX ID is an online cinema ticket booking service or application company made by PT Nusantara Elang Sejahtera. This application provides services such as ordering movie tickets from your favorite cinema, choosing online movies to buy or rent, and buying vouchers to watch streaming movies online. However, the rating accompanied by various reviews does not mean that the wishes or problems of users are fulfilled and stop evaluating in improving services to users. For this reason, sentiment analysis is needed that can classify reviews as user sentiment. In this study, the scraping stage was carried out for collecting application user review data, followed by the text preprocessing stage to process data by selecting data and turning it into more structured data. The data from text preprocessing were word weighted using the Term Frequency – Inverse Document Frequency (TF-IDF) method. Then sentiment classification is carried out using the Support Vector Machine (SVM) algorithm. The best results obtained with the SVM algorithm for sentiment testing of 2 classes with unbalanced data with positive data 274 and negative data 100 using training data values and test data 90%:10%, total data with 374 data, using parameter value  $C = 10$ , cross validation experiment  $K=10$  and use of linear kernel. The results obtained for the average value of 91% accuracy, 94% precision, 83% recall and 86% f-measure.*

*The best results obtained with the SVM algorithm for testing sentiment of 2 classes with balanced data with positive data 150 and negative data 150 using training data values and test data 90%:10%, total data with 300 data, using parameter value  $C = 0.5$ , experiment cross validation  $K=10$  and use of linear kernel. The test results on balanced data get the highest accuracy results, namely 94%, precision 94%, recall 94% and f-measure 94% in the Support Vector Machine method with 2 sentiment classification.*

**Keywords:** *Sentiment Analysis, Scraping, Text Preprocessing, TF-IDF, Support Vector Machine, Google Play Store*

## 1. PENDAHULUAN

Bidang teknologi dan telekomunikasi khususnya teknologi internet berkembang pesat. Berdasarkan data internetworldstat, pengguna Internet di Indonesia mencapai 212,35 juta jiwa pada Maret 2021. Indikator ini menempatkan Indonesia di urutan ketiga dengan pengguna Internet terbanyak di Asia (katadata.co.id). Kebanyakan orang menggunakan internet di semua bidang kehidupan mereka. Salah satunya menggunakan internet untuk mengakses jejaring sosial seperti google, instagram, whatsapp, facebook, telegram. Banyak pengguna yang mengunggah postingan ke media sosial dalam bentuk foto, video, serta teks. Semua yang diunggah akan menjadi data dalam jumlah besar. Data ini dapat diolah dan dianalisis untuk memberikan manfaat di masa depan.

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. Aplikasi ini pada situs layanan Google Play Store telah diunduh sebanyak 10 juta lebih dengan rating 4,5 dan 427 ribu lebih ulasan. Rating yang disertai dengan berbagai ulasan bukan berarti bahwa keinginan atau masalah pengguna terpenuhi dan berhenti untuk melakukan evaluasi dalam peningkatan pelayanan kepada pengguna. Hal ini dikarenakan terdapat beberapa pengguna yang memberikan rating yang tinggi namun ulasan yang diberikan merupakan review negatif. Salah satu kelemahan dari TIX ID adalah seringnya pelanggan mengalami kegagalan dalam melakukan pengisian ulang (top up), sehingga menyebabkan terganggunya penggunaan sistem aplikasi ini. Selain itu beberapa kasus seringnya pengguna juga mengalami pengembalian dana, dan cukup banyak menyebabkan pengguna gagal bayar

secara daring akibat loading yang terlalu lama.

Untuk memudahkan para pihak terkait dalam mengumpulkan informasi tentang kepuasan aplikasi, diperlukan analisis sentimen untuk mengkategorikan data ulasan pengguna aplikasi berdasarkan peringkat dan ulasan di situs web layanan Google Play Store. Analisis emosi atau opinion mining adalah studi komputasi tentang pendapat, perasaan, dan emosi orang melalui entitas dan atribut yang mereka ungkapkan dalam teks (Liu, 2012). Melalui analisis sentimen, perusahaan dapat menemukan wawasan tentang reaksi dan sikap kelompok atau individu terhadap topik diskusi dalam konteks keseluruhan dokumen (Alwasi'a, 2020). Informasi yang diperoleh dari proses ini dapat menjadi dasar bagi perusahaan untuk melakukan inovasi dan perbaikan.

Beberapa penelitian yang telah dilakukan mengenai analisis sentimen pada ulasan aplikasi, antara lain : analisis sentimen Zoom Cloud Meetings di Play Store Menggunakan Naïve Bayes dan Support Vector Machine oleh Herlinawati, dkk (2020). Metode Support Vector Machine yang digunakan menghasilkan nilai akurasi sebesar 81.22%. Penelitian yang serupa telah dilakukan Wahyudi, dkk (2021) tentang Analisis sentimen pada review aplikasi Grab di Google Play Store Menggunakan Support Vector Machine. Disimpulkan bahwa algoritma Support Vector Machine menghasilkan nilai akurasi 85.54%.

Berkaitan dengan hal tersebut, penelitian ini akan menganalisis ulasan aplikasi mobile atau android TIX ID di Indonesia pada situs google play store dengan objek penelitian adalah data ulasan diambil dari bagian ulasan pada situs google playstore review pengguna untuk aplikasi TIX ID. Dalam penelitian ini, penulis akan menerapkan metode klasifikasi Support Vector Machine. Penelitian ini diharapkan dapat digunakan PT Nusantara Elang Sejahtera menjadi landasan dalam meningkatkan mutu dan layanan.

**2. LANDASAN KEPUSTAKAAN**

**2.1. Web Scraping**

Menurut Josi, Abdillah dan Suryayusra (2014), web scraping atau screen scraping dalam bidang data mining artinya adalah suatu usaha untuk memahami pola dan tren semantik dari sejumlah besar data yang diperoleh, tidak dapat diklasifikasikan. Web scraping hanya berfokus pada cara mengambil data dengan mengambil dan mengekstrak data dari berbagai ukuran data.

**2.2. Text Mining**

Penambangan teks adalah tentang memperoleh informasi yang berguna dari sekumpulan dokumen. Menurut Purbo (2019), text mining juga dapat diartikan sebagai proses mengekstraksi, menemukan, memvisualisasikan (menyajikan) dan mengevaluasi pengetahuan dari sejumlah besar dokumen teks.

**2.3. Pembobotan Kata TF-IDF**

Pembobotan kata (*term weighting*) adalah proses pemberian bobot kata ke dokumen. Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan term dalam dokumen. Frekuensi kemunculan (*term frequency*) menunjukkan sejauh mana suatu term merepresentasikan isi dari suatu dokumen. Semakin sering istilah tersebut muncul dalam dokumen, semakin tinggi nilai kesesuaiannya. Bobot dapat mencerminkan pentingnya kata-kata dalam dokumen. Yunus (2020) menjelaskan bahwa pembobotan ini digunakan oleh algoritma machine learning untuk klasifikasi dokumen.

**2.4. Analisis Sentimen**

Menurut Rozi, dkk (2012), analisis sentimen atau opinion mining adalah proses memahami, mengekstrak, dan mengolah data tekstual secara otomatis untuk memperoleh informasi emosional yang terkandung dalam kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecenderungan seseorang tentang suatu masalah atau objek. Tidak masalah jika orang tersebut memiliki opini negatif atau positif.

**2.5. Support Vector Machine**

*Support Vector Machine* (SVM) diperkenalkan oleh Vapnik pada tahun 1992 sebagai teknik klasifikasi yang efisien. SVM memiliki konsep yang lebih lengkap dan jelas

secara matematis dibandingkan teknik klasifikasi sebelum tahun 1990-an (Suyanto, 2019).

**3. METODOLOGI**

Dalam penelitian ini menggunakan metodologi secara terstruktur terdapat 7 tahapan yang telah dipaparkan pada Gambar 1.



Gambar 1. Metodologi Penelitian

Tahap pertama adalah “Studi Literatur” dilakukan dengan mencari sumber-sumber yang mendukung penelitian dan memberikan informasi atau teori dalam penyelesaian penelitian. Tahap selanjutnya adalah “Pengumpulan Data” Pengumpulan data dilakukan dengan cara mengambil data ulasan pengguna aplikasi TIX ID dari kolom ulasan situs layanan *google play store* dengan cara *web scrapping* menggunakan *google play scraper* dengan bahasa pemrograman Python.

Tahap ketiga “Pelabelan Data” adalah melakukan memberikan label positif dan negatif terhadap kalimat ulasan pengguna. Pelabelan data dilakukan secara manual dengan bantuan ahli bahasa Indonesia.

Tahap kelima “Pembobotan Kata” dilakukan pemberian bobot pada *term* dalam teks ulasan. *Term* digunakan untuk mengetahui konteks dari dokumen tersebut dengan cara memberikan indikator. Proses pembobotan menggunakan metode *Term Frequency-Inverse Document Frequency* dengan alat bantu library *Scikit learn*.

Tahap keempat adalah “*Preprocessing Data*” dari data yang telah diperoleh akan

dilakukan proses data *preprocessing* untuk mendapatkan data yang sesuai dengan kebutuhan penelitian. Tahapan preprocessing data yaitu, *cleaning*, *filtering*, dan *stemming*. *Cleaning* yaitu pembersihan kata seperti perubahan huruf besar ke kecil, menghapus *mentions*, *hashtag*, *link*, dan angka. *Stemming* merupakan tahap pengolahan kata menjadi kata dasar dari sebuah kata yang telah memiliki kata imbuhan (baik itu awalan atau akhiran).

Tahap keenam “Klasifikasi Sentimen” Implementasi metode klasifikasi dengan algoritma *Support Vector Machine* (SVM). Pada tahap ini semua algoritma menggunakan bantuan dari *library scikit learn*.

Tahap terakhir “Penarikan Kesimpulan” Tahapan penarikan kesimpulan akan dilakukan ketika sudah mendapatkan hasil dari klasifikasi beserta nilai evaluasinya.

#### 4. PERANCANGAN

##### 4.1. Perancangan Sistem

Pada bab perancangan berisi tahapan rancangan sistem secara menyeluruh dengan diagram alir sistem, proses *scraping* data, pelabelan dataset, *text preprocessing*, manualisasi pembobotan kata, manualisasi klasifikasi *Support Vector Machine* (SVM) yang ditampilkan pada Gambar 2.



Gambar 2. Alur Perancangan Sistem

##### 4.2. Perancangan Skenario Pengujian

Pada rancangan skenario pengujian sistem akan dilakukan skenario pengujian yang dibentuk menjadi dua bagian, untuk bagian

pertama merupakan skenario dengan membandingkan data latih 50-90% dan data uji 10-50%. Bagian kedua adalah perbandingan menggunakan parameter C (Kompleksitas). Tujuan dilakukan pengujian skenario tersebut untuk mendapatkan hasil agar dapat dianalisis apakah dengan menggunakan perbedaan data latih dan data uji serta parameter C (Kompleksitas) akan mempengaruhi hasil dan dengan data manakah akan mendapatkan hasil akurasi terbaik.

#### 5. PENGUJIAN DAN ANALISIS

Pada pengujian ini dibuat berdasarkan rancangan yang dirancang pada bab Perancangan dan hasil proses penerapan sistem pada bab Implementasi. Pengujian pada penelitian ini menggunakan klasifikasi 2 sentimen dengan total data tidak seimbang yang digunakan 374 data yaitu 274 kelas negatif dan 100 kelas positif serta data seimbang dengan data positif dan negatif sebanyak 150. Pembagian data yang akan digunakan untuk pengujian klasifikasi dengan data latih dan data uji sebesar 90%:10% dan parameter C sebesar 10. Hasil pengujian akan ditampilkan berupa hasil *accuracy* dan evaluasi performa dengan pengujian sebanyak sepuluh kali dan dihitung rata – ratanya untuk setiap skenario.

##### 5.1. Skenario Pertama–Pengujian Perbandingan Data Latih dan Uji

Pada pengujian ini dilakukan dengan dua cara membandingkan jumlah data latih dan data uji dengan data seimbang dan tidak seimbang. Pembagian data bertujuan untuk mengetahui keakuratan data yang diuji dengan perbandingan tersebut dan berapa banyak data yang diprediksi benar. Pengujian dilakukan sebanyak 5 variasi yaitu 90%:10%, 80%:20%, 70%:30%, 60%:40%, 50%:50%. Pengujian akan dilakukan sebanyak 10 kali dan hasil akan dirata-rata. Tabel 1 merupakan hasil pengujian pada klasifikasi 2 sentimen data tidak seimbang dengan data positif sebesar 274 dan data negatif sebesar 100.

Tabel 1. Hasil Rerata *Accuracy* Data Tidak Seimbang Skenario Pertama

Latih:Uji	Rerata Accuracy
90%:10%	91%
80%:20%	90%
70%:30%	90%
60%:40%	89%

50%.50%	88%
---------	-----

Tabel 2 merupakan hasil pengujian pengujian pada klasifikasi 2 sentimen data seimbang dengan data positif dan negatif sebesar 150.

Tabel 2. Hasil Rerata Accuracy Data Seimbang Skenario Pertama

Latih:Uji	Rerata Accuracy
90%.10%	97%
80%.20%	95%
70%.30%	96%
60%.40%	97%
50%.50%	91%

### 5.2. Skenario Kedua-Pengujian Nilai Parameter C (Kompleksitas).

Pengujian nilai C metode Support Vector Machine dilakukan untuk mendapatkan nilai akurasi tertinggi. Nilai C yang digunakan yaitu 0.01, 0.1, 0.5, 1, 10. Pengujian pada nilai C diulang sebanyak 10 kali dan hasil dirata-rata. Nilai akurasi C yang diuji untuk klasifikasi 2 sentimen. Hasil akurasi pada parameter nilai C dengan data tidak seimbang dapat dilihat pada Tabel 3. Hasil akurasi pada parameter nilai C dengan data seimbang dapat dilihat pada Tabel 4.

Tabel 3. Hasil Rerata Accuracy Data Tidak Seimbang Skenario Kedua

Latih:Uji	Rerata Accuracy
0.01	73%
0.1	73%
0.5	85%
1	89%
10	90%

Tabel 4. Hasil Rerata Accuracy Data Tidak Seimbang Skenario Kedua

Latih:Uji	Rerata Accuracy
0.01	48%
0.1	90%
0.5	97%
1	96%
10	96%

### 5.3. Skenario Ketiga-Pengujian dengan K-Fold

Evaluasi performa pada Support Vector Machine dengan nilai K sebesar 10 untuk klasifikasi 2 sentimen data tidak seimbang dengan menggunakan presentase data latih dan data uji 90%:10% dapat dilihat pada Tabel 5.

Tabel 5. Hasil Rerata Accuracy Data Tidak Seimbang Skenario Ketiga

Fold ke-	Accuracy	Precision	Recall	F-measure
1	94%	96%	88%	91%
2	88%	93%	77%	82%
3	97%	98%	94%	96%
4	97%	98%	94%	96%
5	85%	91%	72%	76%
6	91%	94%	83%	87%
7	97%	98%	95%	96%
8	81%	81%	80%	72%
9	87%	92%	77%	81%
10	87%	92%	77%	81%
Rata-rata	91%	94%	83%	86%

Evaluasi performa pada Support Vector Machine dengan nilai K sebesar 10 untuk klasifikasi 2 sentimen data seimbang dengan menggunakan presentase data latih dan data uji 90%:10% dapat dilihat pada Tabel 6.

Tabel 6. Hasil Rerata Accuracy Data Tidak Seimbang Skenario Ketiga

Fold ke-	Accuracy	Precision	Recall	F-measure
1	100%	100%	100%	100%
2	88%	90%	88%	88%
3	92%	93%	92%	92%
4	100%	100%	100%	100%
5	96%	96%	96%	96%
6	88%	91%	88%	88%
7	100%	100%	100%	100%
8	81%	82%	81%	81%
9	88%	89%	88%	88%
10	100%	100%	100%	100%
Rata-rata	94%	94%	94%	94%

## 6. PENUTUP

### 6.1. Kesimpulan

Pada pengujian data tidak seimbang menggunakan parameter pengujian yaitu presentase data latih dan data uji sebesar 90%:10%, nilai kompleksitas (C) sebesar 10 dan cv = 10. Pada pengujian data seimbang menggunakan parameter pengujian yaitu presentase data latih dan data uji 90%:10%, nilai kompleksitas (C) sebesar 0.5 dan cv = 10.

Hasil pengujian pada data tidak seimbang mendapatkan hasil akurasi tertinggi yakni 91%, precision 94%, recall 83% dan f-measure 86% pada metode Support Vector Machine klasifikasi 2 sentimen. Hasil pengujian pada data seimbang mendapatkan hasil akurasi tertinggi yakni 94%, precision 94%, recall 94% dan f-measure 94%

pada metode Support Vector Machine klasifikasi 2 sentimen.

### 6.1. Saran

Menambahkan hasil evaluasi ke dalam bentuk visualisasi data agar dapat dijelaskan secara deskriptif. Dengan hasil pengujian yang didapatkan menggunakan Support Vector Machine bahwa hasil tidak seimbang mendapatkan nilai akurasi 91% yang merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan benar. Sisa dari 9% merupakan rasio prediksi salah. Nilai precision 94% merupakan rasio prediksi benar positif. Dari semua kelas positif sisa 6% nya merupakan kelas negatif. Nilai recall 83% merupakan rasio prediksi keberhasilan model dalam menemukan sebuah informasi positif, sisa 17% merupakan tidak keberhasilan model dalam menemukan sebuah informasi. F-measure 86% mengindikasikan bahwa model klasifikasi mempunyai precision dan recall yang baik. Harapannya dengan menggunakan metode lain, akan dapat meningkatkan nilai hasil evaluasi yang lebih baik.

## 7. DAFTAR PUSTAKA

- Alwasi'a, A., 2020. Analisis Sentimen Pada Review Aplikasi Berita Online Menggunakan Metode Maximum Entropy (Studi Kasus: Review Detik.com pada Google Play 2019) [Skripsi]. Yogyakarta: Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Islam Indonesia.
- Josi, Ahmat, Leon Andretti Abdillah, Suryayusra. "Penerapan Teknik Web Scraping Pada Mesin Pencari Artikel Ilmiah." Jurnal Ilmu Komputer Universitas Bina Darma
- Katadata.com. 2021. Pengguna Internet Indonesia Peringkat ke-3 Terbanyak di Asia. Diakses pada 10 Februari 2022, dari <https://katadata.co.id/search/news/pengguna%2520internet/-/-/-/-/10>
- Liu., B., 2012. Sentiment Analysis and Opinion Mining, s.l:Morgan & Claypool Publisher.
- PT Nusantara Elang Sejahtera. (2013). TIX ID Google.com. Diakses pada 17 Februari 2022, dari <https://play.google.com/store/apps/details?id=id.tix.android&hl=in&gl=US>
- Purbo, Onno W. 2019. Text Mining Analisis MedSos, Kekuatan Brand & Intelejen Di Internet. Yogyakarta: Andi.
- Rozi, Imam Fahrur, Sholeh Hadi Pramono, Erfan Achmad Dahlan. 2012. "Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi." Jurnal EECCIS Universitas Brawijaya.
- Sugiyono. 2009. Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, Dan R&D. Bandung: Alfabeta.
- Suyanto. 2019. Data Mining Untuk Klasifikasi Dan Klasterisasi Data. Bandung: Informatika.
- Vulandari, Retno Tri. 2017. Data Mining Teori Dan Aplikasi Rapidminer. Yogyakarta: Gava Media.
- Wahyudi, Rizki, Gilang Kusumawardana. 2021. Analisis Sentimen pada Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine. Jurnal Informatika Universitas Bina Sarana Informatika.
- Yunus, Muhammad. 2020. "TF-IDF (Term Frequency-Inverse Document Frequency) : Representasi Vector Data Text." Diakses pada 10 Ferburari 2022, dari <https://medium.com/@yunusmuhammad/007/tf-idf-term-frequency-inverse-document-frequency-representasi-vector-data-text-2a4eff56cda>