LAPORAN UJIAN AKHIR
ANALISIS BIG DATA
ANALISIS SENTIMEN ULASAN PENGGUNA PADA PLATFORM STEAM MENGGUNAKAN TF-IDF DAN LOGISTIC REGRESSION

Dosen Pengampu:

Ulfa Siti Nuraini, S.Stat., M.Stat.
Sabrina Amelialevi, S.Kom., M.Kom.

Disusun Oleh:

Valentino Prasetyo Putra	(22031554002)
Azwin Rasyiq Azinar	(22031554010)
Siti Sirotul Azhar	(22031554026)
Joevita Salsabila F.	(22031554031)

PROGRAM STUDI S1 SAINS DATA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI SURABAYA

2025

RINGKASAN

Proyek ini bertujuan menganalisis sentimen ulasan pengguna pada platform Steam sebagai dasar pengambilan keputusan dan rekomendasi game berbasis data. Dataset yang digunakan berasal dari Steam Reviews dengan total 350.000 ulasan yang dipilih secara seimbang antara sentimen positif dan negatif. Representasi teks dilakukan menggunakan metode TF-IDF, sementara proses klasifikasi dibandingkan menggunakan tiga algoritma machine learning, yaitu Logistic Regression, Random Forest, dan Support Vector Machine (SVM). Hasil evaluasi menunjukkan bahwa Logistic Regression merupakan model dengan performa paling stabil dan konsisten, ditunjukkan oleh nilai *accuracy* sebesar 0,80 serta keseimbangan F1-score pada kedua kelas sentimen. Analisis *confusion matrix* dan sampel prediksi mengungkap bahwa model berbasis TF-IDF efektif dalam mengenali sentimen eksplisit, namun masih memiliki keterbatasan dalam memahami konteks ulasan yang bersifat ambigu atau temporal. *Insight* yang dihasilkan menunjukkan bahwa DayZ dan Rust didominasi oleh ulasan positif, yang mencerminkan tingkat kepuasan pengguna yang relatif tinggi. Sementara itu, PAYDAY 2, Grand Theft Auto V, dan Dota 2 memiliki jumlah ulasan negatif yang signifikan, dengan PAYDAY 2 juga menunjukkan jumlah ulasan positif yang tinggi, yang mengindikasikan adanya polarisasi sentimen di kalangan pengguna. Temuan ini menegaskan bahwa analisis sentimen mampu memberikan gambaran objektif mengenai persepsi pengguna serta menjadi dasar yang relevan dalam penyusunan rekomendasi game dan evaluasi kualitas produk berbasis data.

LATAR BELAKANG

Data menjadi aset penting dalam pengambilan keputusan berbasis informasi, khususnya pada era digital yang ditandai dengan pertumbuhan data dalam skala besar [1]. Dalam konteks game digital, ulasan pengguna merefleksikan tingkat kepuasan dan pengalaman penggunaan, yang umumnya dinyatakan melalui sentimen positif atau negatif [2]. Platform distribusi game seperti Steam menyediakan volume ulasan yang besar dan kaya informasi, namun sulit dianalisis secara manual. Oleh karena itu, diperlukan pendekatan otomatis berbasis Natural Language Processing (NLP), khususnya analisis sentimen, untuk mengklasifikasikan ulasan pengguna secara efisien [3]. Proyek ini menerapkan metode TF-IDF sebagai representasi fitur teks dan Logistic Regression sebagai algoritma klasifikasi sentimen. Hasil analisis diharapkan dapat mengungkap pola sentimen pengguna serta menghasilkan insight berbasis data yang dapat dimanfaatkan dalam pengambilan keputusan dan rekomendasi game.

ANALISIS DATA

1. Sumber dan Karakteristik Data

Data yang digunakan berasal dari dataset publik Steam Reviews yang diperoleh melalui platform Kaggle, dengan total 6,4 juta data ulasan [4]. Untuk mengurangi beban komputasi selama proses analisis dan pemodelan, proyek ini menggunakan sebanyak 350.000 data yang dipilih dari keseluruhan dataset. Pemilihan data dilakukan dengan mempertimbangkan keseimbangan distribusi kelas sentimen, sehingga jumlah data sentimen negatif dan positif dibuat seimbang. Sampel data disimpan dalam basis data PostgreSQL dan dideploy menggunakan layanan Supabase, yang mendukung pengelolaan data berskala besar secara aman serta koneksi langsung dengan Python melalui library *psycopg2*. Dataset Steam Reviews terdiri atas beberapa atribut utama yang relevan untuk analisis sentimen, sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Karakteristik Atribut Dataset

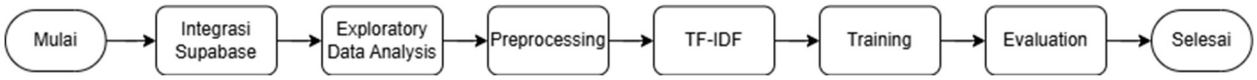
Nama Kolom	Tipe Data	Deskripsi
user id	bigint	Identitas unik pengguna Steam
app_id	text	Identitas unik aplikasi/game
app_name	text	Nama aplikasi/game
review_text	text	Teks ulasan yang diberikan pengguna
review_score	text	kelas sentimen ulasan (-1: negatif, 1: positif)

2. Eksplorasi Data Awal

Tahap eksplorasi data dilakukan untuk memahami karakteristik dataset sebelum memasuki proses pemodelan. Analisis ini mencakup pemeriksaan ukuran data, tipe data, panjang teks ulasan, keberadaan nilai hilang, serta distribusi kelas sentimen. Berdasarkan hasil analisis, panjang teks pada kolom 'review_text' menunjukkan variasi yang cukup besar, dengan panjang minimum 1 karakter, maksimum 8.253 karakter, dan rata-rata 334,36 karakter. Variasi ini menunjukkan heterogenitas teks ulasan, yang berpotensi memengaruhi representasi fitur pada tahap pemodelan.

Pemeriksaan *missing values* menunjukkan adanya data kosong pada kolom 'app_name' sebanyak 12.391 data dan pada kolom 'review_text' sebanyak 387 data. Ulasan kosong tidak mengandung informasi linguistik yang dapat digunakan untuk analisis sentimen, sehingga data tersebut ditangani pada tahap *preprocessing*. Distribusi kelas sentimen pada data terpilih berada dalam kondisi seimbang, dengan masing-masing 175.000 data untuk sentimen negatif dan positif. Keseimbangan ini memberikan kondisi yang menguntungkan bagi proses pelatihan model klasifikasi karena mengurangi risiko dominasi prediksi terhadap salah satu kelas sentimen.

METODE



Gambar 1. Diagram alir proses sistem

Diagram alir proses sistem pada Gambar 1 merepresentasikan alur pengolahan data dan pengembangan model yang dilakukan secara bertahap. Proses dimulai dari integrasi data, dilanjutkan dengan analisis eksploratif, pra pemprosesan, ekstraksi fitur berbasis TF-IDF, pelatihan model, dan diakhiri dengan evaluasi kinerja. Diagram ini bertujuan untuk mempermudah pemahaman alur kerja sistem secara umum. Berikut adalah penjelasan detail pada tiap prosesnya:

1. Preprocessing Data

Proses *preprocessing* dilakukan menggunakan pendekatan *batch processing* dengan ukuran *batch* sebanyak 5.000 data, sehingga diperoleh total 70 *batch* pemrosesan. Pendekatan ini bertujuan untuk menjaga efisiensi penggunaan memori dan waktu komputasi. Tahapan *preprocessing* teks meliputi:

a) *Case Folding*

Seluruh teks ulasan diubah menjadi huruf kecil untuk menghindari perbedaan makna akibat variasi huruf kapital.

b) Penghapusan Karakter Tidak Perlu
Karakter seperti *newline*, *mention*, URL, dan spasi berlebih dihapus untuk menjaga konsistensi teks.

c) Penghapusan Karakter Non-Alfanumerik
Karakter selain huruf dan angka dihilangkan untuk mengurangi noise pada data teks.

d) Normalisasi Slang
Kata-kata slang atau singkatan dinormalisasi menggunakan kamus slang bahasa Inggris yang diperoleh dari dataset publik Kaggle, sehingga istilah tidak baku dapat dikonversi ke dalam bentuk kata yang lebih standar [5].

e) Penghapusan Stopwords
Kata-kata umum dalam bahasa Inggris (stopwords) seperti *the*, *is*, *and*, dihapus karena tidak memiliki kontribusi signifikan terhadap sentimen.

f) Stemming
Proses stemming dilakukan menggunakan *Porter Stemmer* untuk mengubah kata ke bentuk dasarnya, sehingga mengurangi variasi kata yang memiliki makna serupa.

Seluruh tahapan *preprocessing* diterapkan secara konsisten pada setiap *batch* data. Setelah seluruh data digabungkan, diperoleh sebanyak 337.609 data ulasan. Pengurangan jumlah data terjadi akibat penghapusan ulasan kosong dan hasil pembersihan teks. Kolom 'review_text_clean' dihasilkan sebagai representasi teks hasil *preprocessing* dan digunakan pada tahap ekstraksi fitur.

2. Ekstraksi Fitur

Pada tahap ini, teks ulasan yang telah melalui proses *preprocessing* direpresentasikan ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF), yang memberikan bobot lebih tinggi pada kata-kata yang bersifat informatif dalam suatu dokumen relatif terhadap keseluruhan korpus. Ekstraksi fitur dilakukan menggunakan TfidfVectorizer dengan pengaturan parameter sebagai berikut: max_features sebesar 5.000 untuk membatasi jumlah fitur kata dan mengendalikan kompleksitas model; ngram_range (1,2) untuk menangkap informasi unigram dan bigram yang relevan dalam analisis sentimen; min_df sebesar 5 untuk mengabaikan kata atau frasa yang sangat jarang muncul; serta max_df sebesar 0,9 untuk menghilangkan kata-kata yang terlalu umum dan kurang diskriminatif. Hasil dari proses ini berupa matriks fitur TF-IDF berdimensi (n_dokumen × 5.000) yang digunakan sebagai masukan pada tahap pelatihan model klasifikasi.

3. Algoritma Klasifikasi

Pada proyek ini digunakan tiga algoritma klasifikasi yang umum diterapkan pada tugas analisis sentimen berbasis teks, yaitu Logistic Regression, Random Forest, dan Support Vector Machine (SVM). Pemilihan beberapa model bertujuan untuk membandingkan performa pendekatan linear dan ensemble dalam menangani fitur TF-IDF berdimensi tinggi.

a. Logistic Regression
Logistic Regression digunakan sebagai model baseline karena efektif dalam menangani data teks berdimensi tinggi yang direpresentasikan menggunakan TF-IDF. Parameter 'max_iter' ditetapkan sebesar 1.000 untuk memastikan proses optimisasi mencapai konvergensi, sedangkan 'n_jobs' = -

1 digunakan untuk mempercepat pelatihan dengan memanfaatkan seluruh sumber daya prosesor yang tersedia.

b. Random Forest

Random Forest digunakan sebagai pendekatan ensemble berbasis pohon keputusan untuk mengevaluasi kemampuan model non-linear dalam mengklasifikasikan sentimen. Model ini dilatih dengan 'n_estimators' = 100, yang dipilih sebagai kompromi antara kestabilan performa dan efisiensi komputasi.

c. Support Vector Machine (SVM)

Support Vector Machine dengan kernel linear digunakan untuk membangun batas pemisah antara kelas sentimen negatif dan positif. Kernel linear dipilih karena efisien dan sesuai untuk data berdimensi tinggi seperti representasi TF-IDF.

4. Evaluasi Model

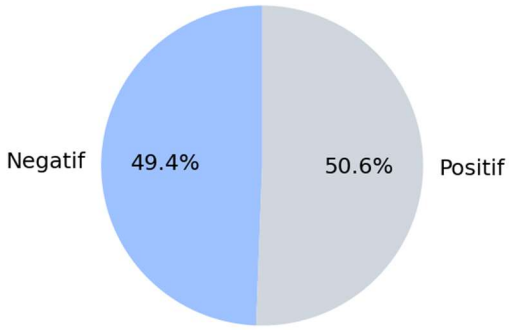
Evaluasi performa model dilakukan menggunakan *confusion matrix* dan *classification report*. *Confusion matrix* digunakan untuk mengamati distribusi prediksi benar dan salah pada masing-masing kelas sentimen, sedangkan *classification report* menyajikan metrik evaluasi berupa *F1-score* dan *accuracy*. Evaluasi dilakukan pada data uji untuk mengukur kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

HASIL DAN EVALUASI

Proyek ini menerapkan analisis sentimen berbasis representasi fitur TF-IDF dengan membandingkan tiga algoritma machine learning. Evaluasi dilakukan melalui analisis distribusi label setelah *preprocessing*, perbandingan kinerja model klasifikasi, serta analisis hasil prediksi secara kuantitatif dan kualitatif.

1. Distribusi Kelas Sentimen Setelah *Preprocessing*

Visualisasi distribusi label sentimen ditunjukkan pada Gambar 2, keseimbangan kelas tetap terjaga setelah tahap *preprocessing* dilakukan. Jumlah data sentimen positif dan negatif berada dalam proporsi yang relatif seimbang, sehingga data yang digunakan pada tahap pelatihan dan pengujian model tidak mengalami bias kelas yang signifikan. Kondisi ini mendukung evaluasi performa model klasifikasi secara lebih objektif.



Gambar 2. Distribusi kelas Setelah *Preprocessing*

2. Evaluasi Kinerja Model Klasifikasi

Evaluasi kinerja model dilakukan dengan membandingkan tiga algoritma klasifikasi, yaitu Logistic Regression, Random Forest, dan SVM, menggunakan metrik *confusion matrix*, *F1-score*, dan *accuracy*. Hasil evaluasi disajikan pada Tabel 2.

Tabel 2. Hasil Evaluasi Kinerja Model

Algoritma	Kelas	Confusion Matrix		F1 Score	Accuracy
Logistic Regression	Positif	36128	5548	0.81	0.80
	Negatif	11606	31121	0.78	
Random Forest	Positif	35778	5898	0.79	0.77
	Negatif	13214	29513	0.76	
SVM	Positif	37938	5812	0.81	0.80
	Negatif	12072	31678	0.78	

3. Analisis Perbandingan Model

Berdasarkan *confusion matrix*, Logistic Regression menunjukkan kinerja paling stabil dan seimbang dalam mengklasifikasikan sentimen positif dan negatif, dengan *accuracy* 0,80 serta *F1-score* 0,81 (positif) dan 0,78 (negatif). Hal ini menunjukkan keseimbangan yang baik antara *precision* dan *recall* pada kedua kelas. SVM memiliki performa yang sebanding dari sisi accuracy (0,80) dan F1-score kelas positif (0,81), namun menghasilkan kesalahan klasifikasi yang sedikit lebih tinggi pada kelas negatif. Sebaliknya, Random Forest menunjukkan performa yang lebih rendah dengan *accuracy* 0,77 dan jumlah kesalahan klasifikasi yang lebih besar, khususnya pada kelas negatif, yang mengindikasikan keterbatasannya dalam menangani representasi TF-IDF yang berdimensi tinggi dan bersifat *sparse*. Secara keseluruhan, evaluasi berdasarkan *confusion matrix* dan metrik kinerja menegaskan bahwa Logistic Regression merupakan model yang paling konsisten dan andal untuk tugas analisis sentimen pada proyek ini.

4. Analisis Sampel Prediksi Model

Untuk memahami perilaku model secara lebih kualitatif, dilakukan analisis terhadap beberapa contoh ulasan dan hasil prediksinya, sebagaimana ditunjukkan pada Tabel 3.

Tabel 3. Hasil Prediksi Tiap Model

Teks Ulasan	Kelas Asli	Prediksi		
		Logistic Regression	Random Forest	SVM
This game is getting lame and why is there so many DLC ???	Negatif	Negatif	Negatif	Positif
Boring after the first few hours.	Negatif	Positif	Positif	Positif
Once every year I play this game start to finish and it's always lovely all over again	Positif	Negatif	Positif	Positif

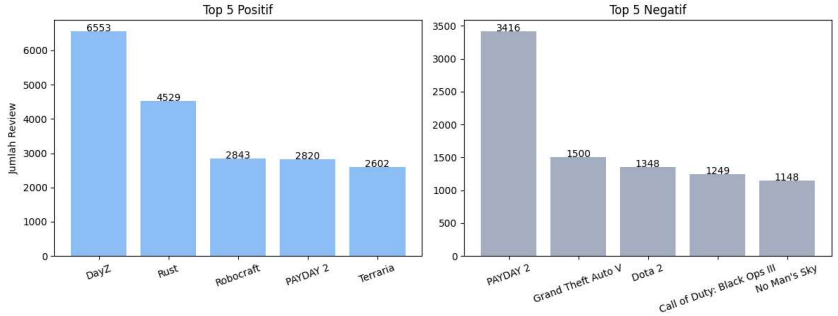
i love this game so much	Positif	Positif	Positif	Positif

Berdasarkan hasil prediksi, seluruh model mampu mengklasifikasikan ulasan dengan ekspresi sentimen yang eksplisit, seperti “i love this game so much”, secara konsisten sebagai sentimen positif, yang menunjukkan efektivitas model dalam mengenali kata-kata dengan polaritas sentimen yang kuat. Namun, kesalahan klasifikasi masih terjadi pada ulasan yang lebih kompleks atau ambigu. Seluruh model salah mengklasifikasikan ulasan “Boring after the first few hours” sebagai sentimen positif, yang mengindikasikan keterbatasan dalam menangkap konteks evaluatif yang bersifat temporal. Selain itu, Logistic Regression keliru memprediksi ulasan positif “Once every year I play this game…” sebagai negatif, sedangkan SVM gagal mengklasifikasikan dengan benar ulasan “This game is getting lame and why is there so many DLC ???”.

Secara keseluruhan, hasil ini menunjukkan bahwa model berbasis TF-IDF cenderung bergantung pada kata kunci tertentu dalam menentukan sentimen, sehingga masih memiliki keterbatasan dalam memahami konteks kalimat yang kompleks atau implisit.

INSIGHT DAN REKOMENDASI

Insight dan rekomendasi disusun berdasarkan hasil analisis sentimen ulasan pengguna pada platform Steam untuk memperoleh pemahaman mengenai persepsi pengguna terhadap kualitas game serta menghasilkan rekomendasi berbasis data. Analisis sentimen dilakukan menggunakan representasi fitur TF-IDF dan diklasifikasikan menggunakan algoritma Logistic Regression, yang berdasarkan hasil evaluasi sebelumnya menunjukkan performa paling stabil dan konsisten dibandingkan model lainnya.



Gambar 3. Visualisasi top 5 game pada tiap sentimen

Hasil visualisasi menunjukkan bahwa DayZ dan Rust didominasi ulasan positif, menandakan tingkat kepuasan pengguna yang tinggi dan konsisten. Sebaliknya, PAYDAY 2, Grand Theft Auto V, dan Dota 2 memiliki jumlah ulasan negatif yang besar. Namun, PAYDAY 2 juga termasuk game dengan ulasan positif tertinggi, yang mengindikasikan adanya polarisasi sentimen akibat basis pengguna yang besar dan tingkat keterlibatan yang tinggi. Temuan ini menunjukkan bahwa analisis sentimen tidak hanya mengidentifikasi dominasi persepsi positif atau negatif, tetapi juga mampu mengungkap dinamika dan kontroversi persepsi pengguna. Oleh karena itu, hasil analisis sentimen pada proyek ini menunjukkan bahwa rekomendasi game tidak dapat ditentukan hanya berdasarkan jumlah ulasan positif, tetapi juga perlu memperhatikan proporsi dan distribusi sentimen untuk menggambarkan persepsi pengguna secara lebih komprehensif. Informasi ini dapat dimanfaatkan sebagai bahan evaluasi awal bagi pengembang dalam mengidentifikasi kekuatan dan potensi permasalahan pada produk berdasarkan pola ulasan pengguna.

DAFTAR PUSTAKA

[1] F. N. Hasan, "Implementasi Sistem Business Intelligence Untuk Data Proyek di Perguruan Tinggi," *Prosiding Seminar Nasional Teknoka*, vol. 4, pp. I1–I10, Nov. 2019, doi: 10.22236/teknoka.v4i1.3943.

[2] E. I. Pantoro, "Harapan Dan Persepsi Konsumen Terhadap Kualitas Layanan Di Kantin Di Universitas Kristen Petra," *Jurnal Hospitality dan Manajemen Jasa*, vol. 5, no. 2, 2017.

[3] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, 2012.

[4] A. Sobkowicz, "Steam Review Dataset (2017)," 2017, *Zenodo*. doi: 10.5281/zenodo.1000885.

[5] M. Rizwan, "social-media-slangs-and-acronyms," Kaggle.