

Analisis Sentimen Ulasan Game Steam Menggunakan Machine Learning dan Model Berbasis Transformer



**Disusun Oleh :
Text Mining - LC09**

Evelyn Caristy Untariady - 2702209496

Malvin Ferdinand Tanzil - 2702208700

**Universitas Bina Nusantara
Kemanggisian
2025**

1. Project Summary

Proyek ini bertujuan untuk melakukan analisis sentimen pada dataset berisi review game yang terdapat di Steam dengan mengkategorikan review dengan label Positif atau Negatif. Proses dimulai dari pembersihan teks menggunakan serangkaian tahapan preprocessing seperti penghapusan simbol (angka dan tautan), normalisasi slang, dan stopwords. Data kemudian direpresentasikan menggunakan TF-IDF serta pemodelan menggunakan dua pendekatan berbeda : algoritma machine learning tradisional (Linear SVC dan Multinomial Naive Bayes) dan model deep learning berbasis transformer (BERT dan DistilBERT).

Hasil evaluasi menggunakan metrik accuracy, precision, recall, dan f1-score menunjukkan bahwa DistilBERT memberikan performa terbaik di antara keempat model dengan akurasi 88.67%.

2. Problem Definition

Masalah utama yang ingin diselesaikan dalam proyek ini adalah menentukan sentimen dari suatu review game di Steam dalam bentuk teks secara otomatis. Masalah ini penting karena platform seperti Steam menerima jutaan ulasan, sehingga sulit bagi pengembang game, analis pasar, atau bahkan calon pembeli untuk menyaring dan memahami pandangan umum komunitas. Model otomatis dapat memberikan wawasan cepat mengenai penerimaan suatu game. Pengembang dan penerbit game adalah pihak yang paling diuntungkan, karena mereka dapat memantau sentimen produk secara real time dan mengidentifikasi area yang perlu perbaikan.

Proyek ini berfokus pada pembangunan model klasifikasi sentimen menggunakan metode text mining. Targetnya adalah menghasilkan model yang akurat, efisien, dan mampu memahami konteks teks dengan baik. Selain itu, proyek ini juga mengevaluasi metode tradisional (TF-IDF + Linear SVC dan Multinomial Naive Bayes) dibandingkan model deep learning berbasis transformer (BERT dan DistilBERT) untuk mengidentifikasi pendekatan paling efektif.

3. Data Collection

Dataset yang digunakan adalah data review game yang berasal dari Kaggle ([Steam Review&Games Dataset](#)) yang datanya diambil (discrepe) dari platform Steam.

Dataset terdiri atas beberapa kolom, yaitu :

- id: Unique identifier untuk masing masing review.
- app_id: Game identifier di Steam.
- content: Text berisi review pengguna.
- author_id: Identifier akun pengguna.
- is_positive: Label yang menunjukkan apakah suatu review positif (1) atau negatif (0).

Secara umum kondisi awal data menunjukkan adanya:

- Teks dengan huruf campuran besar-kecil
- Simbol dan tanda baca
- Tautan URL
- Angka
- Slang (terutama yang sering digunakan di game)
- Adanya duplikasi dan missing values

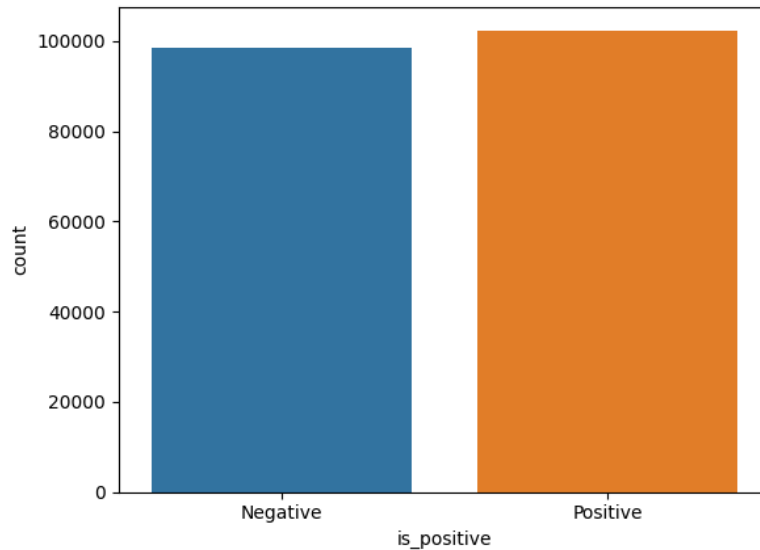
Data mentah ini kemudian dibersihkan dan dipersiapkan sebelum dianalisis.

4. Exploratory Data Analysis (EDA)

EDA dilakukan untuk memahami dataset sebelum melakukan data preprocessing.

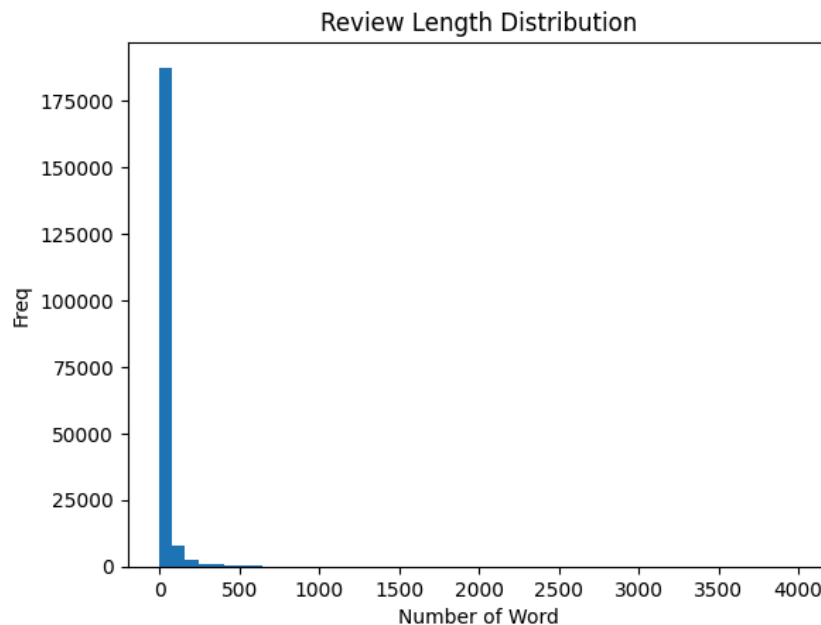
Dari proses EDA terdapat beberapa temuan, yaitu :

- Jumlah Data
Dataset terdiri dari sekitar 201.150 baris, sebelum dilakukan pembersihan duplikasi dan nilai kosong.
- Distribusi Label
Distribusi sentimen (positif dan negatif) cukup seimbang dimana jumlah data dengan sentimen positif sebanyak 102.377 dan dengan sentimen negatif sebanyak 98.346



- Panjang Teks

Ditemukan bahwa rata - rata panjang review sebesar 25.28 kata dimana menunjukkan bahwa teks review yang diberikan oleh pengguna memiliki panjang yang sedang, tidak terlalu panjang maupun pendek.



- Frekuensi Kata (Top Words)

Dari sini ditemukan bahwa kata dominan di tiap sentimen berbeda. Top words untuk sentimen positif berupa : good, best, fun, great, like. Sedangkan top words untuk sentimen negatif berupa : don't , bot, even, time, fix.

- Terdapat Dirty Value

Ditemukan adanya dirty value berupa tautan link dan slang yang biasa digunakan dalam game, yang perlu ditangani pada tahap preprocessing.

5. Data Preprocessing

Tahapan preprocessing diterapkan secara bertahap:

1. Menghapus data duplikat dan data dengan missing value
 - Baris dengan nilai content yang hilang (428 baris) dihapus menggunakan `df.dropna()`.
2. Casefolding
 - Semua teks diubah menjadi huruf kecil (`.lower()`).
3. Menghapus noise (angka, simbol, tanda baca, URL)
 - Karakter non-alfanumerik dan spasi selain spasi tunggal dihapus (`re.sub(r'^\\w\\s|', '', text)`)
 - Tautan URL (`https?:/\\S+|www\\.\\S+`) dihapus
 - Angka (`\\d+`) dihapus
 - Spasi berlebih dihapus
4. Normalisasi slang
 - Beberapa slang umum (lol, lmao, wtf, idk, gg) dinormalisasi menjadi bentuk yang lebih formal atau terstandar (laughing, what the fuck, i dont know, good game).
5. Tokenisasi
 - Memecah kalimat menjadi token menggunakan `nltk.word_tokenize`.
6. Stopwords removal
 - Menghilangkan stopwords general dalam Bahasa Inggris dan stopwords tambahan yang berhubungan dengan game.

Contoh Review :

Sebelum Preprocessing (Raw Text)	Setelah Preprocessing
Alien Swarm: Reactive Drop is basically the same game with more content https://store.steampowered.com/app/563560/Alien_Swarm_Reactive_Drop/	alien swarm reactive drop is basically the same game with more content
even on easy, the last couple missions are exhausting, frustrating and confusing also 1999 gameplay in 2003 lmao	even on easy the last couple missions are exhausting frustrating and confusing also gameplay in laughing

6. Feature Extraction / Text Representation

Terdapat 2 jenis metode text representation yang digunakan, yaitu :

- TF - IDF

TF-IDF digunakan untuk model machine learning tradisional yaitu Multinomial Naive Bayes dan Linear SVC. TF-IDF mengubah teks menjadi vektor numerik yang dimana nilai TF-IDF untuk setiap kata merepresentasikan seberapa penting kata tersebut dalam suatu dokumen relatif terhadap seluruh korpus. Hal ini efektif karena memberikan bobot yang lebih tinggi pada kata-kata yang spesifik dan diskriminatif. Namun TF-IDF memiliki kelemahan yaitu tidak dapat memahami konteks kalimat.

- Transformer Embeddings (BERT dan DistilBERT)

Metode ini dipilih dikarenakan dapat mengatasi kelemahan TF-IDF dengan menghasilkan contextual embeddings, yang mampu memahami makna kata berdasarkan konteks kalimatnya.

7. Modeling

Pemodelan sentimen dilakukan dengan membandingkan empat model berbeda: dua model machine learning (Naive Bayes dan SVM) dan dua model deep learning (Transformer). Data dibagi menjadi Train (80%) dan Test (20%) dengan stratify berdasarkan label is_positive untuk menjaga keseimbangan kelas.

a. Model Machine Learning

Model	Hyperparameter
Multinomial Naive Bayes	-
Linear SVC	-

Kedua model ini memiliki waktu pelatihan yang jauh lebih singkat dibandingkan model transformer. Hasil evaluasi juga menunjukkan bahwa Linear SVC memiliki performa sedikit lebih unggul dibandingkan Multinomial Naive Bayes.

b. Model Transformer

Model	Hyperparameter
DistilBERT	epochs=3, batch_size=16, learning_rate=2e-5
BERT	epochs=3, batch_size=16, learning_rate=2e-5

Model Transformer memerlukan waktu pelatihan yang jauh lebih lama, namun secara signifikan lebih unggul dalam hasil klasifikasi dibandingkan model tradisional. Hasil evaluasi menunjukkan DistilBERT sebagai model terbaik yang menunjukkan keunggulan Contextual Embeddings dalam menangkap semantik teks.

8. Evaluation

Performa model dievaluasi menggunakan metrik Accuracy dan Classification Report (Precision, Recall, F1-Score), yang merupakan metrik standar untuk masalah klasifikasi yang seimbang. Untuk keempat model yang digunakan, berikut adalah hasil evaluasinya :

Model	Recall	Precision	F1-Score	Accuracy
Multinomial Naive Bayes	0.83	0.83	0.83	0.83
Linear SVC	0.84	0.84	0.84	0.84
DistilBERT	0.89	0.89	0.89	0.89
BERT	0.88	0.88	0.88	0.88

Dari hasil evaluasi di atas dapat dilihat bahwa keempat model memiliki performa yang baik dengan akurasi di atas 82%. Untuk model machine learning tradisional, performa model Linear SVC (akurasi sebesar 84%) sedikit lebih unggul dibandingkan model Multinomial Naive Bayes (akurasi sebesar 83%). Meskipun demikian, kinerja model tradisional secara signifikan tertinggal dari model Transformer. Model DistilBERT dan BERT menunjukkan performa terbaik, menegaskan keunggulan contextual

embeddings dalam memahami konteks bahasa review game yang sering kali informal. DistilBERT memimpin dengan akurasi tertinggi sebesar 0.8867 dan skor F1-Score, Precision, dan Recall yang seragam di angka 0.89. Nilai yang konsisten dan tinggi pada semua metrik ini menunjukkan bahwa model tersebut tidak hanya akurat secara keseluruhan, tetapi juga sangat seimbang dalam memprediksi kelas Positif dan Negatif. Sementara itu, model BERT juga menunjukkan performa yang baik dengan akurasi 0.8850, hanya sedikit di bawah DistilBERT. Hal ini menunjukkan betapa pentingnya representasi semantik dalam melakukan sentimen analisis untuk review game.

9. Interpretation / Insights

Insights yang didapatkan dari hasil pemodelan, terutama pada fitur yang dipelajari Linear SVC adalah :

- **Kata Paling Positif (Linear SVC Coefficient):** bad thing, better cs, diamond, well worth, get used, better tf, never gets, classic, dont regret, underrated. Frasa seperti better cs (lebih baik dari Counter Strike) dan well worth adalah indikator kuat sentimen positif.
- **Kata Paling Negatif (Linear SVC Coefficient):** fixtf, dota, savetf, unplayable, boring, cant recommend, age well, moba, dont recommend, unless. Kata-kata seperti unplayable, boring, dan dont recommend merupakan prediktor kuat sentimen negatif. Terdapat juga frasa yang spesifik ke konteks game tertentu, seperti fixtf dan savetf, yang mengindikasikan keluhan terkait game Team Fortress (tf).

Dari hal ini dapat disimpulkan bahwa ulasan negatif sering berpusat pada masalah kinerja (unplayable), gameplay (boring), dan permintaan kepada pengembang (fixtf, valve, bot). Ini menunjukkan bahwa isu teknis dan bot adalah pemicu ketidakpuasan yang besar bagi user.

10. Final Output & Deliverables

Proyek ini menghasilkan output berupa notebook file yang berisi :

- Visualisasi EDA
- Preprocessing Steps
- Model baseline (TF-IDF + Linear SVC dan Multinomial Naive Bayes)
- Model transformer (BERT dan DistilBERT)
- Classification reports dari masing masing model

Hasil dapat diakses di : [evelynuntariady/game_review_sentiment_analysis](https://evelynuntariady.github.io/game_review_sentiment_analysis)

11. Conclusion

Proyek ini telah mencapai tujuan utama yaitu membangun model klasifikasi sentimen pada ulasan game Steam menggunakan pendekatan text mining. Model tradisional memberikan performa baik, namun model transformer menunjukkan peningkatan signifikan, terutama dalam memahami konteks.

Model terbaik adalah DistilBERT, yang memberikan akurasi sekitar 89% dan efisiensi komputasi tinggi. Hasil ini menunjukkan bahwa transformer adalah pilihan unggul untuk analisis sentimen berbasis teks panjang dan variatif seperti ulasan game.

12. Future Improvements

Untuk penelitian kedepannya dapat dilakukan beberapa peningkatan seperti melakukan fin tuning lebih mendalam terhadap hyperparameter seperti learning rate, batch size, max sequence length dan lain sebagainya. Selain itu dapat dilakukan eksperimen dengan model transformer lainnya seperti RoBERTa atau DeBERTa.

13. Reflection

Proses pengerjaan proyek memberikan kami banyak pembelajaran mengenai text mining approach dalam kasus sentimen analisis, terutama dalam membandingkan performa model tradisional dan transformer. Tantangan terbesar yang kami hadapi adalah proses preprocessing text yang dimana harus disesuaikan dengan konteks data dan fine tuning model transformer yang memerlukan sumber daya komputasi besar. Kami juga belajar membagi tugas dan melakukan debugging code secara kolaboratif. Secara keseluruhan, proyek ini membantu kami dalam memperkuat pemahaman tentang aplikasi text mining analisis sentimen di kasus dunia nyata.