

# Analisis Sentimen pada Ulasan Produk Amazon dengan Distillbert

Telkom University

Naufal Arkan Wahib, Hanif Imaduddin, Gabriel Edbert Liandrew

## Pendahuluan

### 1. Latar Belakang

Analisis sentimen telah menjadi salah satu bidang penting dalam pemrosesan bahasa alami yang digunakan untuk memahami opini, emosi, dan sikap pengguna terhadap suatu produk atau layanan. Dengan semakin berkembangnya platform *e-commerce*, seperti Amazon, pengguna memiliki hak untuk memberikan ulasan terhadap produk yang mereka beli. Ulasan ini sangat berharga bagi calon pembeli sebab bisa menjadi bantuan dalam pengambilan keputusan. Ulasan juga berguna bagi pemilik produk sebagai bahan evaluasi kepuasan dari pengguna. Namun, jumlah ulasan yang sangat besar membuat proses analisis manual menjadi tidak efisien dan sulit dilakukan dalam skala besar. Oleh karena itu, dibutuhkan sistem otomatis yang dapat mengklasifikasikan sentimen dari ulasan pelanggan secara cepat dan akurat.

Pendekatan berbasis *machine learning* dan *deep learning* telah banyak diterapkan dalam analisis sentimen. Model tradisional seperti Random Forest, Naïve Bayes, dan *Support Vector Machine* (SVM) telah digunakan secara luas untuk tugas klasifikasi sentimen. Namun, model-model ini memiliki keterbatasan dalam memahami konteks yang kompleks dalam bahasa alami. Sebagai solusi, model berbasis Transformer seperti DistilBERT telah muncul sebagai alternatif yang lebih unggul. DistilBERT adalah versi yang lebih ringan dari *Bidirectional Encoder Representations from Transformers* (BERT) dengan jumlah parameter yang lebih sedikit, tetapi tetap mempertahankan performa yang tinggi dalam berbagai tugas *Natural Language Preprocessing* (NLP). Model ini mampu menangkap hubungan antar kata dalam kalimat dengan lebih baik dibandingkan model tradisional, sehingga lebih efektif dalam analisis sentimen.

Penelitian ini berfokus pada penerapan model DistilBERT untuk mengklasifikasikan ulasan pelanggan dalam kategori positif dan negatif. Selain itu, dalam penelitian ini dilakukan perbandingan dengan model Random Forest untuk mengevaluasi keunggulan metode berbasis Transformer dibandingkan dengan pendekatan *machine learning* konvensional. Dengan menggunakan dataset besar dari platform *e-commerce*, penelitian ini bertujuan untuk mengukur efektivitas dan efisiensi model dalam mengolah teks ulasan pelanggan. Hasil penelitian ini diharapkan dapat memberikan wawasan mengenai model terbaik yang dapat digunakan untuk analisis sentimen secara akurat dan efisien.

### 2. Tujuan

Penelitian ini bertujuan untuk mengembangkan model *machine learning* dan *deep learning* yang mampu melakukan analisis sentimen terhadap ulasan produk di Amazon. Model yang digunakan dalam penelitian ini adalah DistilBERT, sebuah model berbasis *Transformer* yang dioptimalkan untuk tugas NLP, serta Random

Forest, sebuah algoritma berbasis pohon keputusan yang digunakan sebagai pembandingan untuk analisis sentimen berbasis fitur statistik.

### 3. Batas Masalah

Penelitian ini memiliki beberapa batasan sebagai berikut:

- a. Data yang digunakan terbatas pada ulasan produk dari platform Amazon.
- b. Model yang dikembangkan hanya fokus pada analisis sentimen secara umum (positif dan negatif) tanpa mempertimbangkan aspek spesifik dari produk.
- c. Evaluasi performa model dilakukan berdasarkan metrik akurasi dan waktu komputasi, tanpa mempertimbangkan metrik evaluasi lainnya.

## Metodologi

### 1. Metode Analisis

#### a. *Data Collecting*

Tahap pengumpulan data bertujuan untuk menyiapkan data latih dan data uji yang sesuai. Pada dataset yang digunakan, eksplorasi pola diperlukan untuk memahami karakteristik setiap variabel sebelum dilakukan penggabungan data ke dalam set pelatihan dan pengujian.

Dataset yang digunakan dalam penelitian ini mencakup data ulasan dari *website* Amazon. Dataset tersebut disimpan dalam format *Comma-Separated Values* (CSV) dan diolah menggunakan pustaka *pandas* untuk memastikan efisiensi dalam proses pemrosesan data.

#### b. *Data Preprocessing*

Sebelum memasuki model, dataset diproses untuk memudahkan model untuk membaca data yang ada.

1) *Data Cleaning*: Isi teks ulasan dikonversi menjadi huruf non kapital. Proses ini memastikan konsistensi data dalam pemodelan serta sebagai bentuk penyederhanaan untuk mengurangi kompleksitas pemrosesan teks.

2) *Re-labeling Sentiment*: Label kelas yang awalnya terdiri dari 1 (negatif) dan 2 (positif) dikonversi menjadi 0 (negatif) dan 1 (positif).

3) *Data Splitting*: Dalam penelitian ini, dataset dibagi menjadi tiga bagian utama, yaitu data latih, data validasi, dan data uji. Pembagian ini dilakukan untuk memastikan bahwa model yang dikembangkan dapat belajar dari data historis, diuji dalam proses validasi untuk menghindari *overfitting*, dan akhirnya dievaluasi pada data uji untuk mengukur performa model secara objektif.

Secara matematis, pembagian dataset dapat dinyatakan sebagai berikut:

- a) Data latih: 0.6%
- b) Data validasi: 0.2%
- c) Data uji: 20%

di mana  $D_{train}$  merupakan himpunan data latih yang digunakan untuk melatih model,  $D_{val}$  digunakan untuk proses tuning model dan validasi performa selama pelatihan, serta  $D_{test}$  merupakan himpunan data uji yang tidak pernah dilihat oleh model sebelumnya dan digunakan untuk mengukur akurasi akhir model.

4) *Tokenize with DistilBERT*: Data teks ulasan ditokenisasi dengan *DistilBERT*. Teks ulasan dipecah menjadi token-token yang lebih kecil, yaitu menjadi kata

atau subkata. Setiap token yang dihasilkan dikonversi menjadi ID numerik yang sesuai dalam kamus token DistilBERT. Setelah itu, urutan token dipadatkan atau dipotong agar sesuai dengan panjang maksimum yang diterima model.

c. *Modeling*

Dalam penelitian ini, digunakan pendekatan melalui NLP untuk melakukan klasifikasi sentimen. Model yang digunakan adalah DistilBERT dan Random Forest.

1) DistilBERT: model *transformer* yang dikembangkan sebagai versi ringan dari *Bidirectional Encoder Representations from Transformers* (BERT). Model ini menggunakan teknik *knowledge distillation*, yaitu metode kompresi model yang bertujuan untuk mengurangi ukuran dan kompleksitas model tanpa kehilangan performa yang signifikan. Algoritma *machine learning* berbasis ensemble yang terdiri dari banyak *decision tree*. Model ini bekerja dengan cara membangun beberapa pohon keputusan secara acak dan menggabungkan hasil prediksi dari setiap pohon untuk memperoleh hasil klasifikasi yang lebih stabil dan akurat.

2) Random Forest

Dalam penelitian ini, berbagai alat dan teknologi digunakan untuk mendukung proses pengolahan data, pemodelan, serta evaluasi performa model. Pemilihan teknologi ini didasarkan pada efisiensi dan kemampuannya dalam menangani tugas NLP dan *machine learning* secara optimal.

2. Alat dan Teknologi

a. Bahasa Pemrograman

Penelitian ini menggunakan Python sebagai bahasa pemrograman utama. Python dipilih karena memiliki ekosistem yang kaya dengan berbagai pustaka yang mendukung analisis data, NLP, serta pembuatan model *machine learning*.

b. *Framework* dan Pustaka

Beberapa pustaka utama yang digunakan dalam penelitian ini meliputi:

1) *Transformer*: Digunakan untuk memanfaatkan model DistilBERT dalam melakukan tokenisasi dan klasifikasi sentimen.

2) *Scikit-Learn*: Digunakan untuk membangun model Random Forest, serta menyediakan berbagai alat untuk evaluasi model, pra-pemrosesan data, dan pembagian dataset.

3) *Torch (PyTorch)*: Pustaka yang digunakan dalam pelatihan model berbasis *deep learning*.

4) *Pandas dan NumPy*: Digunakan untuk pengolahan data, manipulasi tabel, dan operasi berbasis numerik.

c. *Environment* dan Perangkat Keras

1) *Jupyter Notebook*: Digunakan sebagai *environment* pengembangan utama untuk eksplorasi data, pemodelan, dan analisis hasil

2) GPU (CUDA-enabled): Model *deep learning* dilatih menggunakan GPU yang diakses melalui PyTorch untuk mempercepat proses *training* dan inferensi model.

## Hasil dan Analisis

### 1. Data Collecting

Dataset yang digunakan dalam penelitian ini merupakan kumpulan ulasan pelanggan dari platform Amazon yang mencakup berbagai kategori produk. Data ini terdiri dari beberapa atribut utama, termasuk kelas sentimen, judul ulasan, dan isi ulasan. Sentimen ulasan diklasifikasikan dalam beberapa kategori, yang mencerminkan apakah pelanggan merasa puas, netral, atau kecewa terhadap produk yang mereka beli.

Setiap ulasan memiliki judul singkat yang merangkum opini pelanggan dalam beberapa kata, serta isi ulasan yang lebih rinci, yang berisi pengalaman, kelebihan, atau kekurangan produk yang dirasakan oleh pengguna. Judul ulasan bervariasi, mulai dari ungkapan kepuasan, seperti *"Excellent"* dan *"Great"*, hingga kekecewaan seperti *"Disappointed"*, dan *"Don't waste your money"*. Dataset ini mencakup ratusan ribu entri ulasan dengan variasi judul yang beragam sehingga mencerminkan luasnya cakupan opini pelanggan di berbagai kategori produk. Informasi terkait dataset dapat dilihat pada tabel 1.

Tabel 1. Informasi dataset

Kelas	Judul	Ulasan
2	Stuning even for the non-gamer	This sound track was beautiful! It paints the ...
2	The best soundtrack ever to anything.	I'm reading a lot of reviews saying that this ...
⋮	⋮	⋮
1	compact, but hard to clean	We have a small house, and really wanted two o...
1	what is it saying?	not sure what this book is supposed to be. It ...
2	Makes My Blood Run Red-White-And-Blue	I agree that every American should read this b...

### 2. Data Preprocessing

#### a. Data Cleaning

Proses pembersihan data dilakukan untuk memastikan konsistensi teks ulasan dengan mengonversinya menjadi huruf non kapital. Hal ini membantu dalam mengurangi kompleksitas pemrosesan teks. Hasil dari konversi bisa dilihat pada tabel 2.

Tabel 2. Perbandingan data sebelum dan sesudah *data cleaning*

Sebelum	Sesudah
The high chair looks great when it first comes out of the box but it is all down hill after that. It is impossible to keep clean. The finish is flaking off, after less than 6 months of use. It is not worth the struggle to keep up.	the high chair looks great when it first comes out of the box but it is all down hill after that. it is impossible to keep clean. the finish is flaking off, after less than 6 months of use. it is not worth the struggle to keep up.
I have used this highchair for 2 kids now and finally decided to sell it because I did not like it. It's a beautiful chair and looks great in the kitchen.	i have used this highchair for 2 kids now and finally decided to sell it because i did not like it. it's a beautiful chair and looks great in the kitchen.

b. *Re-Labelling Sentiment*

Pada tahap ini, label sentimen yang awalnya berupa angka 1 (negatif) dan 2 (positif) diubah menjadi 0 (negatif) dan 1 (positif). Perubahan ini dilakukan agar lebih sesuai dengan konvensi umum dalam pemodelan NLP dengan label biner sering kali dikodekan dengan 0 untuk negatif dan 1 untuk positif. Hasil dari konversi bisa dilihat pada tabel 3.

Tabel 3. Perbandingan data sebelum dan sesudah *re-labelling sentiment*

Sebelum	Sesudah
1	0
2	1

c. *Data Splitting*

Dataset dibagi menjadi tiga bagian utama untuk memastikan model dapat berlatih, divalidasi, dan diuji dengan data yang sesuai. Sebaran dari subset data bisa dilihat di tabel 4.

Tabel 4. Informasi pembagian data

Subset	Proporsi	Jumlah Data
Data Latih	0.6%	21.600
Data Validasi	0.2%	7.200
Data Uji	20%	720.000

Dataset yang digunakan memiliki 3.600.000 sampel. Data ini terlalu besar untuk diproses secara langsung oleh DistilBERT tanpa memerlukan komputasi yang sangat tinggi. DistilBERT juga merupakan *pre-trained model* sehingga tidak memerlukan data yang besar untuk melakukan pelatihan. Oleh karena itu, dilakukan pengambilan sampel untuk data latih dan validasi. Untuk data uji, seluruhnya digunakan sebagai evaluasi objektif yang memastikan model bekerja dengan baik.

d. *Tokenize with DistilBERT*

Proses tokenisasi dilakukan menggunakan **DistilBERT**, di mana teks ulasan diubah menjadi serangkaian token yang lebih kecil dan dikonversi menjadi ID numerik yang dapat dipahami oleh model. Sampel hasil proses tokenisasi bisa dilihat pada tabel 5.

Tabel 5. Sampel hasil *tokenize*

Token	ID Token
[CLS]	101
i	1045
love	2293
this	2023
cd	3722
.	1012
saw	2387
them	2068
in	1999
person	2711
...	...
SEP	102

### 3. Modeling

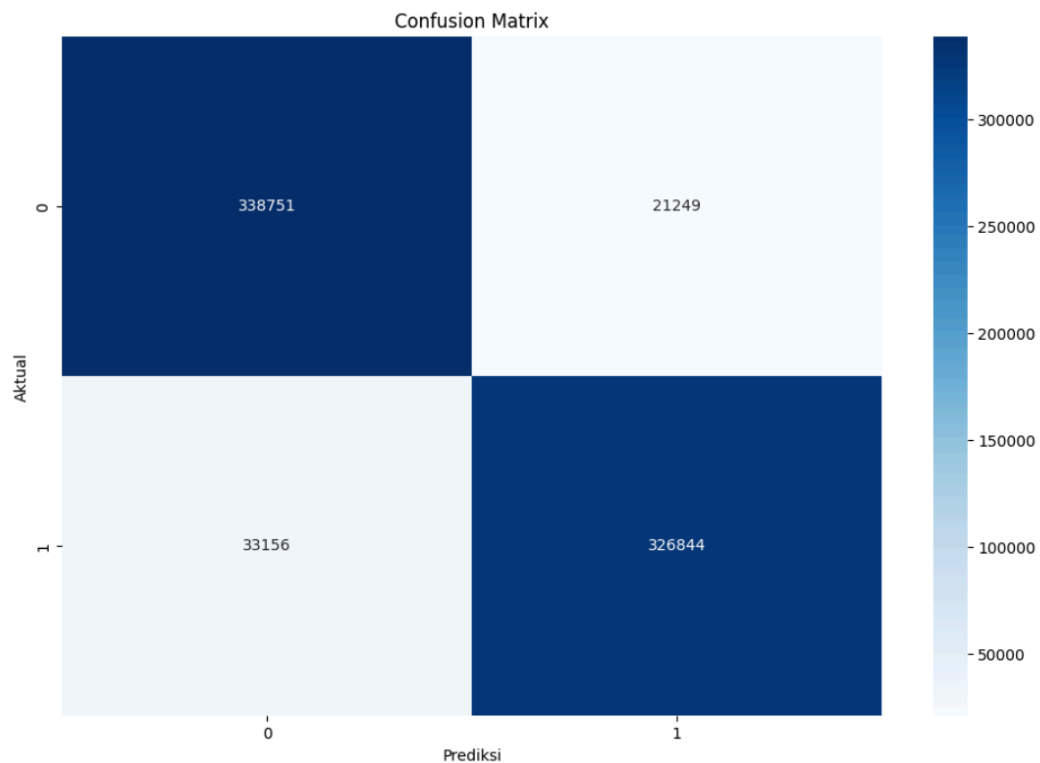
Dari tabel 6, dapat dilihat bahwa DistilBERT memiliki akurasi yang jauh lebih tinggi (92,44%) dibandingkan Random Forest (61,05%). Hal ini menunjukkan bahwa DistilBERT mampu memahami konteks dalam teks lebih baik dibandingkan model berbasis pohon keputusan seperti Random Forest.

Tabel 6. Perbandingan evaluasi model

Model	Akurasi	Waktu Testing
RandomForest	0.6105	5 jam
DistilBERT	<b>0.9244</b>	<b>3 jam</b>

Salah satu faktor utama yang menyebabkan perbedaan ini adalah cara kedua model dalam memahami teks. Random Forest bekerja berdasarkan ekstraksi fitur eksplisit, seperti jumlah kemunculan kata atau skor TF-IDF, tanpa mempertimbangkan konteks kalimat secara keseluruhan. Sebaliknya, DistilBERT menggunakan arsitektur *Transformer*, yang mampu menangkap hubungan antar kata dalam berbagai konteks, sehingga memberikan pemahaman yang lebih baik terhadap makna teks.

Selain itu, DistilBERT memiliki kemampuan generalisasi yang lebih baik dibandingkan Random Forest. Model berbasis pohon keputusan seperti Random Forest cenderung bekerja dengan baik pada dataset kecil hingga menengah, tetapi kesulitan dalam menangani teks dengan struktur yang lebih kompleks. Sebaliknya, DistilBERT telah melalui tahap pre-training pada jumlah data yang sangat besar, sehingga memiliki pemahaman yang lebih luas terhadap bahasa alami. Hal ini membuatnya lebih unggul dalam menganalisis teks yang memiliki makna tersirat atau frasa yang sulit dipahami oleh model berbasis aturan sederhana.



Gambar 1. *Confusion matrix* data uji model DistilBERT

Dari segi efisiensi waktu, DistilBERT juga lebih cepat dalam proses pengujian meskipun memiliki arsitektur yang lebih kompleks. Dalam eksperimen ini, DistilBERT hanya membutuhkan 3 jam untuk melakukan pengujian, sementara Random Forest memerlukan 5 jam untuk jumlah data yang sama. Meskipun Random Forest sering dianggap sebagai model yang ringan, waktu inferensi menjadi lebih lambat pada skala data yang besar karena banyaknya pohon keputusan yang harus dievaluasi dalam proses klasifikasi. Sementara itu, DistilBERT yang telah dioptimalkan melalui teknik *knowledge distillation* mampu melakukan inferensi lebih cepat tanpa kehilangan banyak akurasi dibandingkan model *Transformer* lainnya. Selain itu, DistilBERT memiliki kemampuan lebih baik dalam memahami teks yang kompleks dan memiliki banyak variasi. Model seperti Random Forest hanya mengandalkan frekuensi kata atau fitur numerik lainnya tanpa memahami hubungan antar kata dalam suatu kalimat. Sebaliknya, DistilBERT menggunakan mekanisme self-attention, yang memungkinkan model untuk menyesuaikan bobot perhatian pada setiap kata dalam kalimat sesuai dengan konteksnya. Hal ini menjadikannya lebih unggul dalam menangani kalimat dengan struktur bahasa yang beragam dan makna yang bergantung pada susunan kata.

## Penutup

### 1. Kesimpulan

Penerapan model deep learning dalam analisis sentimen sangat penting untuk memahami opini pengguna dalam skala besar. Pada penelitian ini, model DistilBERT dibandingkan dengan model Random Forest untuk mengklasifikasikan sentimen ulasan pelanggan. Hasilnya menunjukkan bahwa DistilBERT memiliki akurasi yang jauh lebih tinggi dibandingkan Random Forest, yaitu 92,44% dengan 61,05%. Hal ini

menunjukkan bahwa model berbasis Transformer lebih unggul dalam menangkap konteks dan makna dalam teks dibandingkan model berbasis aturan atau statistik tradisional.

Dari segi efisiensi waktu, DistilBERT juga lebih cepat dalam proses pengujian, hanya membutuhkan 3 jam, sedangkan Random Forest memerlukan 5 jam untuk dataset yang sama. Dengan keunggulan dalam akurasi dan efisiensi, DistilBERT lebih cocok untuk diterapkan dalam sistem analisis sentimen berskala besar.

## 2. Rekomendasi

Berdasarkan hasil analisis yang diperoleh, maka dapat dituliskan beberapa rekomendasi sebagaimana terdapat pada Tabel 7.

Tabel 7. Rekomendasi

Aspek	Rekomendasi
Peningkatan Akurasi	Menggunakan teknik <i>fine-tuning</i> pada DistilBERT untuk menyesuaikan model dengan karakteristik dataset.
Efisiensi Waktu	Mengoptimalkan komputasi dengan menggunakan batch processing dan mixed precision training agar proses inferensi lebih cepat.
Perbandingan Model	Membandingkan performa DistilBERT dengan model berbasis Transformer lainnya seperti BERT atau RoBERTa.

## Daftar Pustaka

Jacob, J. (2023). *Sentiment Analysis with DistilBERT*. Diakses pada 23 Maret 2025, dari <https://jacobj215.github.io/Sentiment-Analysis-with-DistilBERT/>

Universitas Gadjah Mada. (2022). *Analisis Sentimen Ulasan Produk E-Commerce Menggunakan Model Transformer*. Diakses pada 24 Maret 2025, dari <https://etd.repository.ugm.ac.id/penelitian/detail/230575>