

## **ANALISIS SENTIMEN ULASAN PRODUK MENGGUNAKAN LARGE LANGUAGE MODELS : STUDI KASUS PADA SHOPEE**

*Ribka Amelia Yunita<sup>a</sup>, Evta Indra<sup>a</sup>*

*<sup>a</sup>Universitas Prima Indonesia*

### **ABSTRACT**

Ketika e-commerce berkembang pesat, jumlah ulasan produk di situs seperti Shopee telah meningkat. Ulasan ini menjadi sumber informasi yang bagus bagi penjual yang ingin lebih memahami perspektif pelanggan dan meningkatkan strategi pemasaran mereka. Karena jumlah dan kompleksitas bahasa yang digunakan untuk menilai, analisis manual menjadi tidak efektif. Penelitian ini bertujuan mengevaluasi sentimen ulasan produk Shopee dengan menggunakan model Large Language Models (LLM), terutama model Gemini 1.5 Pro yang disesuaikan dengan bahasa pelanggan Shopee. Teknik ini mencakup pengumpulan data melalui ekstraksi data menggunakan web, praproses data, penyempurnaan model dan evaluasi kinerja model menggunakan skor F1, akurasi, presisi dan recall. Temuan penelitian menunjukkan bahwa model yang disarankan dapat mengklasifikasi sentimen ulasan ke dalam tiga kategori: positif, negatif, dan netral, dengan rentang akurasi dari 66,67% hingga 85,71%.

**Keywords:** Analisis Sentimen, Shopee, LLMs, Gemini 1.5 Pro.

### **1. INTRODUCTION**

Karena pesatnya perkembangan internet dan teknologi, platform e-commerce telah berkembang menjadi komponen penting dalam operasi perdagangan internasional (Alamin et al., 2023). Berbagai macam barang elektronik dan kebutuhan sehari-hari membuat Shopee menjadi salah satu toko online terbesar di Indonesia (Muktafin et al., 2020). Salah satu komponen utama yang memudahkan komunikasi antara pelanggan dan pedagang Shopee adalah sistem ulasan produk. Ulasan pelanggan adalah cara yang bagus bagi penjual untuk meningkatkan barang dan layanan mereka serta bagi calon pelanggan untuk mempelajari lebih lanjut tentang produk dan layanan mereka (Pradhana et al., 2024). Meskipun sistem ulasan ini memiliki banyak manfaat, analisis manusia terhadap banyaknya evaluasi ulasan terkadang sulit dilakukan (Kamal & Ratnasari, 2021). Akibatnya, metode yang lebih canggih diperlukan untuk analisis sentimen otomatis untuk evaluasi produk (Arrasyid et al., 2024).

Sulit bagi penjual dan administrator platform e-commerce untuk memahami dan menangani ratusan hingga ribuan ulasan yang dibuat setiap hari. Kadang-kadang, penilaian ini mencakup perasaan yang tidak jelas, seperti “baik” atau “buruk”, selain menggunakan bahasa yang lebih bernuansa, seperti ironi, ambiguitas, atau penggunaan emotikon (Sherly Christina, 2019). Selain itu, interpretasi perasaan menjadi lebih sulit karena bahasa konsumen yang beragam. Termasuk bahasa lokal, bahasa gaul, dan ekspresi artistik lainnya (Alita et al., 2019). Hal ini membuat sulit untuk menangani variasi ini menggunakan analisis sentimen manusia atau dengan metode yang lebih konvensional seperti pembelajaran mesin berbasis fitur statistik (Arrasyid et al., 2024).

Dalam beberapa tahun terakhir, Large Language Models (LLM) seperti Gemini 1.5 Pro, GPT-3 dan GPT-4 telah muncul, menunjukkan kemajuan pesat dalam teknologi Natural Language Processing (NLP) (Liu et al., 2024). Menurut (Ahuja et al., 2024), dalam bahasa yang lebih kompleks, model ini dapat memahami sikap yang

disimpulkan secara langsung atau tidak langsung dalam evaluasi produk. Large Language Models (LLM) seperti BERT dan GPT telah menjadi populer untuk analisis sentimen (Limbong et al., 2022). Penelitian sebelumnya telah menunjukkan bahwa LLM dapat memprediksi sentimen dengan lebih akurat daripada metode yang lebih tradisional seperti Naive Bayes atau Support Vector Machines. Ini terutama berlaku untuk teks yang lebih beragam dan bebas seperti yang ditemukan dalam ulasan platform e-commerce (Muktafin et al., 2020).

Bahkan dengan kemajuan teknologi, masih ada masalah besar yang harus diatasi saat menggunakannya untuk melihat ulasan produk di situs seperti Shopee. Ini terutama karena pengguna Shopee berbicara berbagai bahasa, dan model harus disesuaikan agar sesuai dengan penggunaannya. Selain itu, pelanggan sering menggunakan emotikon dan peringkat bintang untuk menyatakan pendapat mereka.

Penelitian ini bertujuan untuk membuat sistem analisis sentimen untuk penilaian produk Shopee menggunakan Gemini 1.5 Pro. Tujuannya adalah untuk meningkatkan analisis sentimen dengan menyesuaikan model Gemini 1.5 Pro dengan bahasa yang digunakan pelanggan Shopee (Limbong et al., 2022). Selain itu, penelitian ini adalah untuk mengatasi masalah dalam menentukan pendapat tentang fitur produk tertentu, seperti harga atau kualitas material, yang sering menjadi fokus ulasan di platform e-commerce (Ari Bangsa et al., 2020). Dengan menggunakan metode ini, penjual dapat mengetahui lebih banyak tentang kepuasan dan preferensi pelanggan, dan pembeli dapat membuat pilihan yang lebih baik tentang apa yang akan dibeli berkat analisis yang lebih akurat dan dapat dipercaya (Roumeliotis et al., 2024).

## **2. LITERATURE REVIEW**

Analisis sentimen digunakan dalam penambahan data data untuk mengkategorikan persepsi konsumen terhadap suatu produk atau layanan. Beberapa makalah penelitian sebelumnya telah menunjukkan bahwa berbagai metode untuk menganalisis sentimen efektif. Sebelumnya telah dilakukan penelitian tentang dampak prapemrosesan teks pada akurasi model Naive Bayes untuk kategorisasi sentimen ulasan IMDb. Temuan menunjukkan bahwa persiapan data, termasuk stemming dan penghapusan stopwords, memengaruhi akurasi model. Kumpulan data praproses memiliki tingkat akurasi tertinggi (83,99%), diikuti oleh stemming (82,36%) dan penghapusan stopwords (83,63%). Namun, model Naive Bayes tidak berhasil menangani teks yang mengalami perubahan struktural akibat prapemrosesan; ini terutama berlaku untuk stemming, yang membuat model sulit memahami kata-kata yang telah direduksi ke bentuk paling dasar (Hakim, 2021).

Dalam penelitian serupa (Dewi et al., 2022), teknik Naive Bayes digunakan untuk mengevaluasi tanggapan konsumen terhadap produk masker di Tokopedia. Dalam mengklasifikasikan penilaian produk ke dalam kategori positif dan negatif, metode Naive Bayes memiliki tingkat akurasi 88%, presisi 61%, dan recall 65%, menurut temuan penelitian. Karena keterbatasan kumpulan data yang digunakan, hasil penelitian tersebut tidak dapat memvalidasi semua evaluasi produk di Tokopedia. Selain itu, metode Naive Bayes kesulitan menemukan perasaan sarkastik atau ambigu dalam struktur frasa yang kompleks.

Penelitian yang dilakukan oleh (Dyah Fritama et al., 2023), menyelidiki kekurangan algoritma K-Nearest Neighbor (KNN) dalam analisis sentimen evaluasi produk perawatan jerawat pada platform Female Daily. Penelitian tersebut menemukan bahwa akurasi kategorisasi berdasarkan merek produk berbeda, dengan Whitelab mencapai tingkat akurasi tertinggi (97%), diikuti oleh Skin Game (81%) dan ERHA (75%). Karena penelitian ini

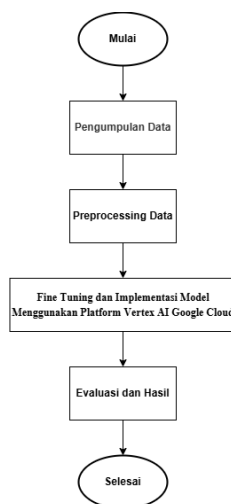
hanya melihat kumpulan data dari tiga merek, hal ini kurang umum terjadi pada berbagai solusi perawatan jerawat yang tersedia secara umum. Ketergantungan metode KNN yang besar pada pilihan parameter analisis K dan mengabaikan hubungan semantik antarkata dapat berdampak pada akurasi sentimen.

Selain itu, penelitian yang dilakukan oleh (Rivaldi et al., 2024) menggunakan metode Natural Language Processing (NLP) untuk memeriksa perasaan pelanggan tentang evaluasi produk di Zalika Store 88 Shopee. Dengan tingkat akurasi sebesar 87%, penelitian ini menunjukkan bahwa metode NLP dapat secara otomatis mengkategorikan evaluasi pelanggan ke dalam kategori baik, negatif, atau netral. Hasil analisis menunjukkan bahwa sebagian besar evaluasi bersifat positif dengan hanya beberapa yang netral atau tidak positif. Dengan menggunakan analisis sentimen yang didasarkan pada pemrosesan NLP, penelitian ini menunjukkan bagaimana pemilik toko dapat meningkatkan strategi bisnis mereka dan lebih memahami sikap klien.

### 3. METHODS

Metode yang digunakan dalam penelitian ini adalah metode deskriptif dan metode analitis (Muzaki et al., 2024). Penelitian ini merupakan jenis penelitian kuantitatif yang bertujuan untuk mengukur, mengolah, dan mengevaluasi data ulasan produk secara statistik untuk memberikan hasil yang dapat dipercaya (Amaldin & Ina Ratnasari, 2024). Sementara metode analitis digunakan untuk mempelajari secara menyeluruh bagaimana LLM dapat digunakan untuk analisis sentimen dan seberapa efektifnya untuk memberikan informasi untuk pengambilan keputusan di e-commerce, metode deskriptif digunakan untuk menggambarkan karakteristik ulasan produk (Rivaldi et al., 2024).

Setiap tahap dari penelitian ini direncanakan untuk memastikan bahwa model yang digunakan dapat memberikan analisis sentimen yang akurat dan dapat diandalkan, karena penelitian ini menggunakan sejumlah langkah metodis untuk menyelidiki dan mengevaluasi sikap konsumen terhadap barang-barang Shopee. Tiap tahapan tersebut dijelaskan pada diagram alir sebagai berikut:



Gambar 1. Diagram Alir Tahapan Penelitian

#### Pengumpulan Data

Langkah pertama dalam proses pengumpulan data penelitian ini adalah menentukan target produk. Dengan menekankan industri kecantikan, khususnya perawatan kulit, penulis memilih kategori produk yang akan dibahas.

Setelah menentukan kategori, penulis mengumpulkan ulasan produk dari platform Shopee menggunakan scraping web, yaitu console.apify.com. Fungsionalitas web tersebut memungkinkan proses scraping dilakukan secara otomatis, sehingga ulasan yang dibutuhkan dapat dikumpulkan secara otomatis.

#### **Preprocessing Data**

Pra-pemrosesan adalah tahap selanjutnya setelah pengumpulan data selesai. Tahap ini menghilangkan elemen yang tidak diperlukan, seperti data duplikat, spasi ganda, dan karakter khusus (Rokhman et al., 2020). Menurut buku panduan pelatihan model Gemini 1.5 Pro, data yang telah dibersihkan terlebih dahulu disimpan dalam format CSV sebelum kemudian diubah ke format JSONL untuk memudahkan langkah pemrosesan berikutnya.

#### **Fine Tuning dan Implementasi Model**

Penulis kemudian menggunakan dataset untuk memperbaiki model tersebut. Metode ini seringkali memerlukan penyesuaian parameter model untuk meningkatkan akurasi analisis sentimen. Pengujian data yang sudah memiliki label sentimen (positif, negatif, dan netral) seringkali diperlukan untuk meningkatkan model analisis sentimen (Rodriguez Inserte et al., 2024). Penulis mengkategorikan semua sentimen ulasan menggunakan model setelah selesai, bukti bahwa hasil klasifikasi menunjukkan emosi ulasan baik, negatif, atau netral.

#### **Evaluasi dan Hasil**

Penulis melanjutkan ke tahap penilaian setelah mengumpulkan hasil kategorisasi. Untuk mengukur efektivitas model yang digunakan penulis dan memastikan keakuratan hasil analisis, evaluasi ini menggunakan metrik seperti akurasi, presisi, recall, dan skor F1; selain itu, evaluasi ini melakukan pengujian manual untuk mengukur kinerja model, serta visualisasi distribusi sentimen. Dengan menyajikan hasil penelitian sebagai statistik deskriptif, penulis akan dapat memberikan ringkasan distribusi sentimen dalam evaluasi produk. Tabel dan grafik yang mudah dipahami akan digunakan untuk menampilkan informasi.

## **4. RESULTS**

#### **Pengumpulan Dataset**

Bagian dari kumpulan data yang digunakan dalam penelitian ini adalah sepuluh ulasan produk perawatan kulit dari sepuluh gerai resmi Shopee. Console.apify.com mengumpulkan data secara otomatis melalui webscraping yang mengekstraksi informasi langsung dari situs produk e-commerce. Gambar infografi sberikut menunjukkan barang-barang dari setiap toko:



Gambar 2. Infografis Toko dan Produk

Toko-toko ini dipilih berdasarkan ketersediaan dan popularitas produk perawatan kulit yang berbeda. Ulasan untuk sepuluh jenis produk perawatan kulit yang paling umum dikumpulkan dari masing-masing toko, yaitu:

1. **Night Cream** – Krim malam yang membantu regenerasi kulit selama tidur.
2. **Moisturizer** – Pelembap yang menjaga kelembapan dan elastisitas kulit.
3. **Mugwort Mask** – Masker berbahan Mugwort, dikenal efektif untuk menenangkan kulit sensitif dan berjerawat.
4. **Facial Wash** – Pembersih wajah untuk menghilangkan kotoran, minyak, dan sisa makeup.
5. **Peeling** – Produk eksfoliasi yang mengangkat sel kulit mati.
6. **Acne Serum** – Serum khusus untuk merawat kulit berjerawat.
7. **Toner** – Cairan penyegar yang menyeimbangkan pH kulit setelah mencuci muka.
8. **Serum** – Produk dengan bahan aktif tinggi untuk mengatasi masalah kulit tertentu.
9. **Sunscreen** – Produk yang melindungi kulit dari sinar UV.
10. **Day Cream** – Krim pagi yang melembapkan dan melindungi kulit dari faktor lingkungan.

Jika toko tidak menjual salah satu produk di atas, produk tersebut akan digantikan dengan produk perawatan kulit yang serupa. Penelitian menunjukkan bahwa pelanggan Indonesia menggunakan merek *skincare* ini paling sering. Oleh karena itu, pemilihan merek *skincare* ini didasarkan pada fakta ini. Sebagai contoh, data dari ZAP Beauty Index 2020 menunjukkan bahwa 92,4% perempuan Indonesia menggunakan pembersih wajah, diikuti oleh pelembab (62,6%), toner (58,3%) dan serum (50,7%) (ZAP Beauty Index, 2020).

Didapat dataset ulasan produk sebesar 50.000 setelah menggunakan console.apify.com untuk mengumpulkan data. Informasi yang dikumpulkan termasuk akun pengguna, komentar pengguna, peringkat layanan kurir, dan peringkat layanan toko. Tujuan penjelasan ini adalah untuk mengetahui bagaimana konsumen melihat berbagai produk perawatan kulit yang tersedia. Gambar sampel data hasil web-scraping dari aplikasi Shopee ditunjukkan sebagai berikut:

| author_username | comment                                                                                               | detailed_rating/product_quality |
|-----------------|-------------------------------------------------------------------------------------------------------|---------------------------------|
| aldurra_16      | Tekstur:lembut, tidak lengket dan harum aroma tehDaya serap:cepat keringEfektivitas:melembutkan       | 5                               |
| z****a          | Barang diterima dengan baik,original,pengiriman juga cepat, pelayanan sangat memuaskan. Akan t        | 5                               |
| n****h          | Daya serap:Cepat banget nyerapnyaTekstur:Nggak terlalu cair, tapi juga nggak terlalu kentalEfektiv    | 5                               |
| b****6          | Jdi aku beli di tanggal 13, dan Aku pake sampe tanggal 26 dan udah habis/sisa dikit kaya di video, U  | 5                               |
| erliningsih05   | aaaa suka banget sama acne serum scarlett ini 😊 jerawat aku super mendingan banget &cpt kerin         | 5                               |
| g****           | Daya serap:gampang banget nyerep ga bandel juga bleber kemana manaTekstur:cairfor the first tir       | 5                               |
| s****h          | Tekstur:Cair, wangiDaya serap:Cepat menyerapEfektivitas:Kulit jadi lebih freshCocok banget buat n     | 5                               |
| n****3          | Yeeyyyy paket udah datang. Maaf baru kasih penilaian 🙏 moga² serumnya cocok di muka aku 🙏             | 5                               |
| d****a          | Cepet bgt nyampinya. Semoga cocok yaa. Klo cocok pasti aku beli trus. 🙏 Adminnya juga baik, rar       | 5                               |
| vivlr05         | Serumnya ga ada bau yg aneh atau wangi kaya parfum gitu. Teksturnya cair, ga lengket, dan cepat n     | 5                               |
| meiwu_          | Pengiriman cepat! Pesan tgl 11, dikirim tgl 12, sampe tgl 13. Pengemasan baik, diplastikin, dikardusi | 5                               |
| uskhusa         | Alhamdulillah cocok pake serum nya scarlett ini, udah botol kedua ini. Packing super aman, bubble     | 5                               |
| febiolamartina  | Udah hari ketiga pake serumnya, serumnya enakEngga bikin semacem rasa nyeri gitu, padahal di m        | 5                               |
| a****           | pengemasannya super rapi dan cute sekali, suka banget deh pokoknya, temiat dan terencana seki         | 5                               |
| megaw07_        | produk super super original sudah aku cek via web nya , semoga aku cock dgn serum nya , pengema       | 5                               |
| s****9          | Ini serum pertama yang aku coba sejauh ini ngga ada penolakan dari kulit aku, alhamdulillah berun     | 5                               |
| j****t          | Petjah abiss 🙏🙏 Safety banget pakingnya...Padahal cuman beli satu produk, dan puas banget san         | 5                               |

Gambar 3. Sampel Hasil Web-scraping

## Preprocessing Data

Setelah pengumpulan data selesai, data harus diproses untuk menghilangkan elemen yang tidak diperlukan. Untuk memastikan bahwa data analisis sentimen berkualitas tinggi dan bebas dari materi yang dapat mengubah hasil penelitian (Hakim, 2021).

Prosedur praproses dijalankan menggunakan ahasa pemrograman Python di platform Google Colab karena lingkungan berbasis cloudnya, yang memiliki banyak pustaka pemrosesan teks dan kemampuan GPU (Fahrezi & Permana, 2022). Beberapa metode praproses termasuk normalisasi teks, analisis tokenisasi, dan penghapusan stopword yang tidak menambah sentimen.

Dalam analisis teks, praproses data sangat penting karena dapat meningkatkan keakuratan interpretasi sentimen model (Sanger, 2011). Pembersihan data yang baik juga dapat meningkatkan kinerja analisis sentimen hingga 20% dibandingkan dengan data mentah tanpa praproses (Zhang et al., 2020). Kumpulan data menjadi lebih bersih setelah pra-pemrosesan ini dan siap untuk analisis sentimen dan ekstraksi fitur. Untuk mengurangi gangguan, karakter khusus seperti tanda baca, simbol, dan emoji dihilangkan. Selain itu spasi ganda atau pengikisan teks dihilangkan untuk mempercepat proses. Hanya peringkat produk dan konten ulasan yang disimpan dalam metadata, seperti akun dan waktu dikecualikan dari ulasan duplikat untuk menghindari bias. Pada tahap ini, kumpulan data yang lebih baik dibuat dan disiapkan untuk analisis lebih lanjut. Di bawah ini adalah contoh kode persiapan data yang digunakan Google Colab:

```

# Pembersihan Data
# Pilih kolom 'comment' dan 'detailed_rating/product_quality'
data = data[['comment', 'detailed_rating/product_quality']]

# Hapus baris dengan nilai yang hilang
data = data.dropna()

# Hapus karakter non-alfanumerik dari kolom 'comment'
data['comment'] = data['comment'].apply(lambda x: re.sub(r'^a-zA-Z0-9\s', '', str(x)))

# Ubah teks menjadi huruf kecil pada kolom 'comment'
data['comment'] = data['comment'].apply(lambda x: x.lower())

# Hapus spasi ekstra pada kolom 'comment'
data['comment'] = data['comment'].apply(lambda x: re.sub(r'\s+', ' ', x))

# Simpan data yang sudah dibersihkan
data.to_csv('shopee_reviews_cleaned.csv', index=False)

# Tampilkan data yang sudah dibersihkan
print(data) # atau print(data.head()) untuk menampilkan 5 baris pertama

```

Gambar 4. Sampel Perintah Pre-Processing Data

Seperti yang ditunjukkan pada Gambar 4, proses pertama pemrosesan data dari ulasan Shopee tentang produk perawatan kulit menggunakan kode Python adalah memilih kolom yang sesuai, yaitu “ evaluasi rinci/kualitas produk” dan “ulasan pelanggan”. Selain itu, data yang tidak ada digunakan untuk memastikan hanya data yang

akurat yang digunakan. Setelah data ulasan diproses, mereka disimpan dalam file CSV bernama “shopee\_reviews\_cleaned.csv”.

Setelah disimpan, dataset diubah menjadi JSONL dari format CSV, sehingga model Gemini 1.5 Pro dapat menggunakannya. Kemudian, dataset dibagi menjadi tiga kelompok menggunakan Python dan dijalankan di Google Colab. Data pelatihan mewakili 90% dataset, data validasi mewakili 9,7%, dan data uji mewakili 0,3% dari dataset. Pola diidentifikasi dengan data pelatihan, data validasi memungkinkan parameter disesuaikan, dan data uji menilai kinerja model. Dataset disimpan di cloud untuk pembersihan dan pemartisian. Contoh perintah pembagian data adalah sebagai berikut:

```
train_data, temp_data = train_test_split(dataset, test_size=0.10, random_state=42)

if len(temp_data) > 1:
    test_size_adjusted = min(0.99, 0.03) # Adjust to a value less than 1 and ensures at least one for training
    # Pisahkan 9.7% untuk testing dan 0.3% untuk evaluasi dari 10% yang tersisa
    test_data, eval_data = train_test_split(temp_data, test_size=test_size_adjusted, random_state=42)
else:
    # Handle the case when temp_data has too few samples
    print("Warning: Not enough samples in temp_data for further splitting. Using temp_data for testing and skipping evaluation.")
    test_data = temp_data
    eval_data = pd.DataFrame() # Create an empty DataFrame for eval_data

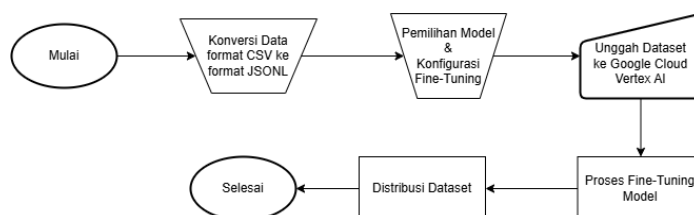
print(f"Train size: {len(train_data)} ({len(train_data)/len(dataset)*100:.2f}%)")
print(f"Test size: {len(test_data)} ({len(test_data)/len(dataset)*100:.2f}%)")
print(f"Eval size: {len(eval_data)} ({len(eval_data)/len(dataset)*100:.2f}%)")

# Simpan hasil dan download
train_data.to_csv('train_data.csv', index=False)
files.download('train_data.csv')
test_data.to_csv('test_data.csv', index=False)
files.download('test_data.csv')
```

Gambar 5. Perintah Pembagian Dataset

### Fine Tuning dan Implementasi Model

Setelah praproses, tahap berikutnya adalah memilih dan mengoptimalkan model untuk analisis sentimen ulasan produk. Google Vertes AI platform cloud yang memudahkan penerapan, penyempurnaan, dan pelatihan model, digunakan dalam proyek ini. Untuk meningkatkan relevansi dan akurasi klasifikasi sentimen, model tersebut disesuaikan agar sesuai dengan kumpulan data tertentu. Berikut adalah alur kerja penyempurnaan model:

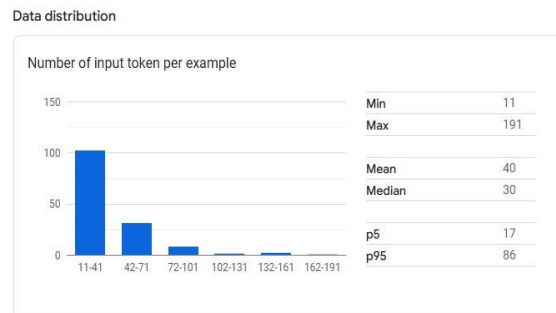


Gambar 6. Alur Proses Fine-Tuning

Selama penyempurnaan, beberapa variabel diubah untuk mendapatkan hasil terbaik. Proses ini disebut “sentiment-classification”, dan model dasar yang digunakan adalah “gemini-1.5-pro-002”. Lokasi untuk server yang digunakan adalah “Asia-Southeast 1 (Singapura). Model dilatih selama 30 kali untuk memaksimalkan proses belajar, dengan Learning Rate Multiplier sebesar 1 untuk mengontrol kecepatan proses penyempurnaan. Adapter Size ditetapkan 16 agar parameter disesuaikan tanpa mengubah struktur model. Pengaturan ini, akan meningkatkan akurasi analisis sentimen. Setelah mengubah set data ke format JSONL dan mengunggahnya ke model, perlu dilakukan penyesuaian tambahan.

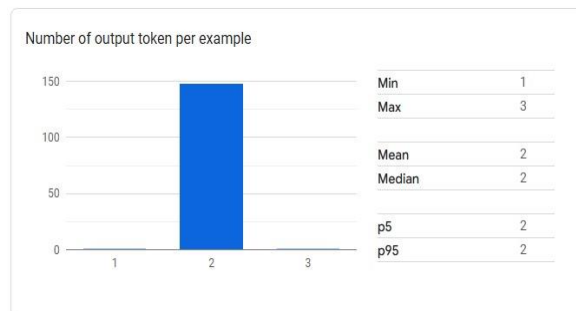
Setelah dataset disesuaikan, data dibagi menjadi 150 data validasi, 4.850 data pengujian, dan 45.000 data pelatihan. Pembagian ini menggunakan metode pembelajaran mesin konvensional untuk memastikan kinerja

model yang optimal. Biasanya, 10% dari dataset digunakan untuk pengujian, 10% untuk validasi, dan 80% untuk pelatihan. Pengujian menilai kinerja model menggunakan data baru, model dilatih, validasi membantu menghindari overfitting dan meningkatkan hyperparameter (Shah, 2017). Setelah itu, distribusi data diperiksa berdasarkan jumlah token input, token output, dan pesan untuk memastikan keseimbangan data dan meningkatkan ketepatan model. Berikut adalah hasil dari distribusi data tentang jumlah token input untuk setiap kasus:



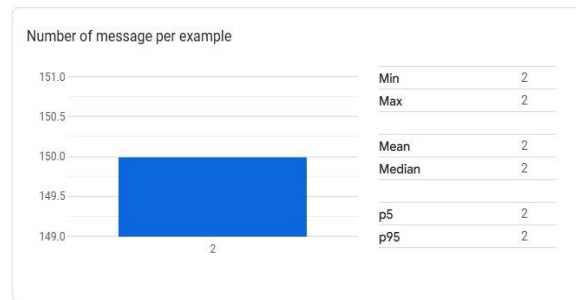
Gambar 7. Distribusi Data: Jumlah Token Input Per Contoh

Setelah penyempurnaan, distribusi jumlah token input per contoh pada Gambar 7 menunjukkan bahwa kebanyakan data terdiri dari teks singkat. Jumlah token berkisar dari minimum 11 hingga maksimum 191, dengan rata-rata 40 dan median 30. Sebagian besar data berada sekitar persentil p5 dan 95, yang masing-masing adalah 17 dan 86 token. Sekitar 40 contoh termasuk dalam kisaran 42-71 token, sementara kebanyakan contoh memiliki 11-41 token dengan lebih dari 100 contoh. Beberapa token melampaui 100, dan hampir tidak ada yang mencapai 191. Distribusi ini menunjukkan bahwa kumpulan data yang telah disempurnakan memiliki beberapa contoh teks panjang dan sebagian besar terdiri dari teks pendek hingga sedang. Gambar berikutnya adalah distribusi data jumlah token output untuk setiap contoh:



Gambar 8. Distribusi Data: Jumlah Token Output per Contoh

Distribusi data jumlah token keluaran per contoh setelah penyempurnaan digambarkan dalam Gambar 8. Model menghasilkan keluaran yang pendek dan konsisten, dengan jumlah token bervariasi antara 1 dan 3. Rata-rata, median dan persentil p5 dan p95 semuanya adalah 2, yang menunjukkan panjang keluaran yang seragam. Kebanyakan contoh memiliki 2 token, dengan variasi minimal dan hampir tidak ada keluaran yang berisi 1 atau 3 token. Ini menunjukkan bahwa model tersebut paling cocok untuk tugas klasifikasi atau pelabelan teks yang keluaran pendek, seperti prediksi sentimen atau klasifikasi berdasarkan label pendek. Model tersebut tidak ditunjukkan untuk pembuatan teks panjang karena terbatasnya token yang dihasilkan. Selanjutnya gambar di bawah ini merupakan ilustrasi jumlah pesan per contoh:



Gambar 9. Distribusi Data: Jumlah Pesan per Contoh

Gambar 9 mengilustrasikan bahwa setiap contoh dalam kumpulan data memiliki jumlah pesan yang tetap setelah penyempurnaan; semua statistik minimum, maksimum, rata-rata, median, dan persen hingga p5 dan p95 adalah 2, yang menunjukkan tidak ada variasi. Visualisasi juga hanya menunjukkan satu batang pada angka 2, yang menunjukkan struktur kumpulan data yang seragam, kemungkinan berupa tanya jawab sederhana atau percakapan dua arah; jika model akan digunakan untuk percakapan yang lebih beragam, kumpulan data perlu lebih bervariasi. Setelah distribusi analisis data dan penyempurnaan selesai, model diterapkan pada kumpulan data yang disiapkan, dan langkah berikutnya adalah pengujian dengan data baru untuk memastikan hasilnya sesuai dengan yang diantisipasi.

## 5. Reporting Research Results

Setelah proses penyempurnaan model selesai, kinerja model dievaluasi dalam tiga tahap: (1) pengukuran seperti akurasi, presisi, recall, dan skor F1; (2) evaluasi visualisasi distribusi sentimen; dan (3) pengujian manual untuk mengetahui kemampuan model untuk membuat prediksi yang akurat.

Dengan menggunakan Google Colab, evaluasi model dilakukan dengan mengukur metrik utama dan menampilkan distribusi sentimen. Berikut adalah contoh perintah yang digunakan untuk menghitung metrik dan menampilkan evaluasi:

```
# Hitung metrik evaluasi
accuracy = accuracy_score(true_labels, predicted_labels)
precision = precision_score(true_labels, predicted_labels, average="weighted")
recall = recall_score(true_labels, predicted_labels, average="weighted")
f1 = f1_score(true_labels, predicted_labels, average="weighted")

# Tampilkan hasil evaluasi
print("\n✅ Evaluation Results:")
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1 Score: {f1:.4f}")

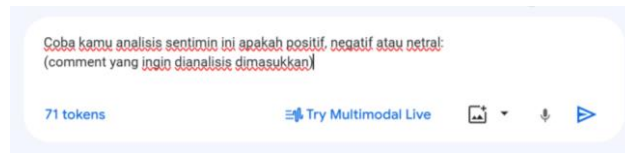
# Hitung distribusi sentimen untuk visualisasi
distribution = pd.Series(predicted_labels).value_counts(normalize=True) * 100

distribution.plot(kind="bar", title="Distribusi Sentimen", xlabel="Sentimen", ylabel="Persentase (%)")
```

Gambar 10. Perintah Evaluasi Metrik dan Visualisasi

Tahap evaluasi manual dilakukan dengan alat obrolan Google Cloud Vertex AI. Ini menentukan seberapa baik model dapat membuat prediksi yang akurat. Pada titik ini, berbagai masukan tinjauan yang tidak ada dalam

set data pelatihan digunakan untuk mengevaluasi respons model terhadap teks yang baru. Berikut merupakan perintah yang digunakan untuk melakukan evaluasi manual:



Gambar 11. Perintah Evaluasi Manual

Berikut merupakan nilai dari evaluasi dan hasil model yang dilakukan:

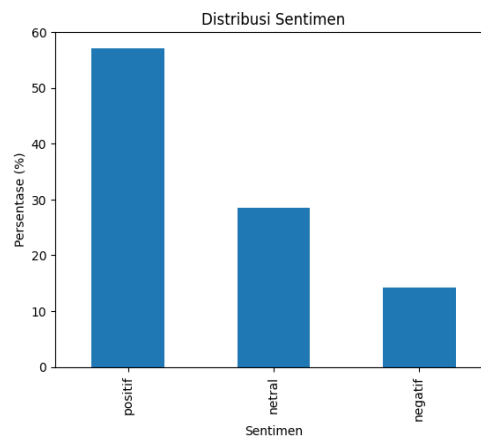
### Evaluasi dan Hasil Uji

Penilaian pertama mengukur metrik sentimen utama: Akurasi, Presisi, Recall dan Skor F1. Tabel berikut menampilkan hasil evaluasi model:

| Metrik Evaluasi | Accuracy | Precision | Recall | F1 Score |
|-----------------|----------|-----------|--------|----------|
| Hasil Evaluasi  | 0.8571   | 0.8929    | 0.8571 | 0.8531   |

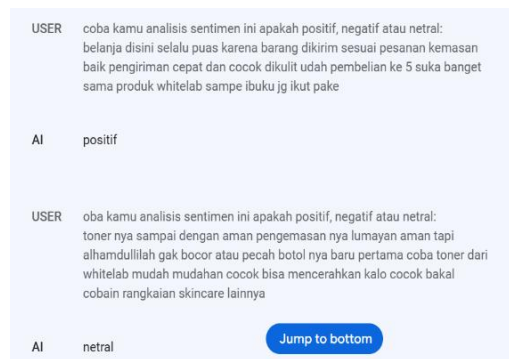
Tabel 1. Evaluasi Metrik

Hasil evaluasi model menunjukkan akurasi sebesar 85,71% dan presisi sebesar 89,29%, yang menunjukkan bahwa model mampu mengklasifikasikan data dengan benar dalam sebagian besar kasus. Recall sebesar 85,71% menunjukkan bahwa model berhasil mengenali sebagian besar sampel positif. Skor F1 sebesar 85,31% digunakan untuk menilai keseimbangan antara presisi dan recall, menunjukkan bahwa model melakukannya dengan baik untuk menyeimbangkan kedua metrik tersebut. Selanjutnya merupakan hasil evaluasi visualisasi distribusi sentimen ditampilkan dalam grafik berikut:



Gambar 12. Evaluasi Visualisasi Sentimen

Secara keseluruhan, hasil visualisasi distribusi sentimen menunjukkan bahwa produk atau layanan yang dianalisis memiliki persepsi positif yang dominan, dengan lebih banyak ulasan yang mendukung daripada ulasan netral atau negatif. Sebagian besar ulasan bersifat positif (lebih dari 50%), yang menunjukkan kepuasan pengguna terhadap produk atau layanan; sentimen netral (sekitar 30%), yang biasanya menunjukkan ulasan informatif tanpa ekspresi emosional yang kuat; dan sentimen negatif (kurang dari 30%). Selanjutnya merupakan evaluasi teks manual melalui Google Cloud Vertex AI dapat dilihat sebagai berikut:



Gambar 13. Evaluasi Teks Manual

Hasil evaluasi manual yang dilakukan melalui fitur obrolan Google Cloud Vertex AI menunjukkan kemampuan model untuk mengkategorikan sentimen berdasarkan konteks kalimat pengguna. Dalam tinjauan pertama, model menemukan sentimen positif berdasarkan kepuasan pengguna terhadap pengiriman, kualitas produk, dan pengalaman pembelian berulang; dalam tinjauan kedua, model menemukan sentimen netral yang sesuai karena tinjauan tersebut bersifat tidak pasti atau ambigu.

## 6. DISCUSSION

Studi ini menggunakan Large Language Model (LLM), khususnya model Gemini 1.5-Prop, untuk menganalisis sentimen ulasan produk di Shopee. Model ini disempurnakan menggunakan kumpulan data ulasan produk Shopee untuk meningkatkan akurasi klasifikasi sentimen dan memberi tahu penjual tentang tren pelanggan. Namun, masalah tetap termasuk sarkasme, ambiguitas, dan efek emotikon dan peringkat dalam ulasan. Studi ini menunjukkan bahwa LLM memiliki keunggulan dalam memahami konteks bahasa yang lebih kompleks, dibandingkan dengan studi sebelumnya menggunakan Naive Bayes dan KNN. Namun, ada masalah untuk menjaga integritas kelas sentimen dalam kumpulan data. Selain itu, hasil analisis menunjukkan bahwa model tersebut memiliki kecenderungan lebih besar untuk mengklasifikasikan ulasan sebagai positif daripada negatif atau netral.

Keterbatasan penelitian ini termasuk bias distribusi sentimen dan fakta bahwa data hanya berasal dari Shopee. Untuk penelitian yang lebih lanjut, disarankan untuk memperluas dataset ke platform lain, menyelidiki model yang lebih kompleks, dan meningkatkan metode pra-prosesan untuk menangani variasi bahasa yang lebih besar. Temuan penelitian ini dapat membantu penjual memahami perasaan pelanggan, membuat respons yang lebih baik terhadap ulasan, dan membuat strategi pemasaran yang lebih baik. Oleh karena itu, diharapkan bahwa model yang dikembangkan akan memberikan manfaat yang lebih besar bagi sektor e-commerce dan meningkatkan akurasi analisis sentimen.

## 7. CONCLUSION

Hasil penelitian ini menunjukkan bahwa Fine-Tuning Large Language Models (LLM) untuk analisis sentimen ulasan produk di Shopee telah berhasil. Menurut hasil evaluasi, model yang dikembangkan dapat mengklasifikasikan ulasan sentimen ke dalam kategori positif, negatif, dan netral. Hasil menunjukkan bahwa model dapat mengidentifikasi sentimen dengan cukup baik, terutama mendeteksi sentimen positif dan negatif.

Namun, ada beberapa masalah, seperti bias model terhadap kelas tertentu dan akurasi yang rendah pada beberapa uji coba.

## 8. REFERENCES

- Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., Hada, R., Jain, P., Ahmed, M., Bali, K., & Sitaram, S. (2024). MEGEVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, 1, 2598–2637. <https://doi.org/10.18653/v1/2024.naacl-long.143>
- Alamin, Z., Missouri, R., Sutriawan, S., Fathir, F., & Khairunnas, K. (2023). Perkembangan E-commerce: Analisis Dominasi Shopee sebagai Primadona Marketplace di Indonesia. *J-ESA (Jurnal Ekonomi Syariah)*, 6(2), 120–131. <https://doi.org/10.52266/jesa.v6i2.2484>
- Alita, D., Priyanta, S., & Rokhman, N. (2019). Analysis of Emoticon and Sarcasm Effect on Sentiment Analysis of Indonesian Language on Twitter. *Journal of Information Systems Engineering and Business Intelligence*, 5(2), 100. <https://doi.org/10.20473/jisebi.5.2.100-109>
- Amaldin, A. D. P., & Ina Ratnasari. (2024). Pengaruh Ulasan Positif dan Ulasan Negatif Produk di E-Commerce Shopee terhadap Keputusan Pembelian Konsumen (Studi Kasus di Fakultas Ekonomi dan Bisnis Universitas Singaperbangsa Karawang). *Madani : Jurnal Ilmiah Multidisipline*, 2(6), 256–86–92.
- Ari Bangsa, M. T., Priyanta, S., & Suyanto, Y. (2020). Aspect-Based Sentiment Analysis of Online Marketplace Reviews Using Convolutional Neural Network. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 14(2), 123. <https://doi.org/10.22146/ijccs.51646>
- Arrasyid, R. M., Putera, D. E., & Yusuf, A. Y. P. (2024). Analisis Sentimen Review Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing. *Jurnal Tekno Kompak*, 18(2), 319. <https://doi.org/10.33365/jtk.v18i2.3813>
- Dewi, E., Mulyani, S., Hidayatuloh Taopik, A., & Agustiawan, T. (2022). Analisis Sentimen Ulasan Produk Pada Top Brand Produk Masker Di Tokopedia Menggunakan Naive Bayes. *Jurnal Eksplora Informatika*, 1(1), 289–292. <https://orcid.org/0000-0002-8366-6697>
- Dyah Fritama, S., Raymond Ramadhan, Y., & Andayani Komara, M. (2023). Analisis Sentimen Review Produk Acne Spot Treatment di Female Daily Menggunakan Algoritma K-Nearest Neighbor. *Media Online*, 4(1), 134–143. <https://doi.org/10.30865/klik.v4i1.1070>
- Fahrezi, M. F., & Permana, A. A. (2022). Sentimen Analisis Opini Masyarakat Pada Sosial Media Twitter Terhadap Organisasi Aksi Cepat Tanggap Menggunakan Naïve Bayes Classifier. *JT: Jurnal Teknik*, 11(02), 113–121. <http://jurnal.umt.ac.id/index.php/jt/index>
- Hakim, B. (2021). Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning. *JBASE - Journal of Business and Audit Information Systems*, 4(2), 16–22. <https://doi.org/10.30813/jbase.v4i2.3000>
- Kamal, W. W., & Ratnasari, C. I. (2021). Analisis Sentimen Ulasan Produk: Kajian Pustaka. *Automata*, 2(1), 1–4.

- Limbong, J. J. A., Sembiring, I., & Hartomo, K. D. (2022). Analisis Klasifikasi Sentimen Ulasan pada E-Commerce Shopee Berbasis Word Cloud dengan Metode Naive Bayes dan K-Nearest Neighbor. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 9(2), 347. <https://doi.org/10.25126/jtiik.2022924960>
- Liu, M., Okuhara, T., Dai, Z., Huang, W., Okada, H., Furukawa, E., & Kiuchi, T. (2024). Performance of Advanced Large Language Models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese Medical Licensing Examination: A Comparative Study. *MedRxiv*, 2024.07.09.24310129. <http://medrxiv.org/content/early/2024/07/09/2024.07.09.24310129.abstract>
- Muktafin, E. H., Kusrini, K., & Luthfi, E. T. (2020). Analisis Sentimen pada Ulasan Pembelian Produk di Marketplace Shopee Menggunakan Pendekatan Natural Language Processing. *Jurnal Eksplora Informatika*, 10(1), 32–42. <https://doi.org/10.30864/eksplora.v10i1.390>
- Muzaki, A., Febriana, V., & Cholifah, W. N. (2024). ANALISIS SENTIMEN PADA ULASAN PRODUK DI E-COMMERCE DENGAN METODE NAIVE BAYES. 05(04), 758–765.
- Pradhana, F. R., Musthafa, A., & Fitria, I. (2024). ANALISIS SENTIMEN ULASAN PRODUK DAVIENA DI SHOPEE. November, 1–13.
- Rivaldi, R. C., Wismarini, T. D., Lomba, J. T., & Semarang, J. (2024). Analisis Sentimen Pada Ulasan Produk Dengan Metode Natural Language Processing (NLP) (Studi Kasus Zalika Store 88 Shopee). 17(1), 120–128.
- Rodriguez Inserte, P., Nakhlé, M., Qader, R., Caillaut, G., & Liu, J. (2024). Large Language Model Adaptation for Financial Sentiment Analysis. 1–10. <https://doi.org/10.18653/v1/2023.finnlp-2.1>
- Rokhman, N., Ningtyas, A. M., Salim, M. F., & Santoso, D. B. (2020). Penerapan Sistem Data Cleansing untuk Mencegah dan Menghilangkan Duplikasi Rekam Medis. *Jurnal Pengabdian Kepada Masyarakat (Indonesian Journal of Community Engagement)*, 6(4), 248–252. <https://doi.org/10.22146/jpkm.51073>
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). LLMs in e-commerce: A comparative analysis of GPT and LLaMA models in product review evaluation. *Natural Language Processing Journal*, 6(January), 100056. <https://doi.org/10.1016/j.nlp.2024.100056>
- Sanger, R. F. and J. (2011). The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data. *Critical Sociology*, 37(4), 493–497. <https://doi.org/10.1177/0261018311403863>
- Shah, T. (2017). Train, Validation and Test Sets. *Tarang Shah - Blog*, 2017–2019. <http://tarangshah.com/blog/2017-12-03/train-validation-and-test-sets/>
- Sherly Christina. (2019). Sarcasm in Sentiment Analysis of Indonesian Text: A Literature Review. *Jurnal Teknologi Informasi Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, 13(2), 54–59. <https://doi.org/10.47111/jti.v13i2.255>
- ZAP Beauty Index. (2020). ZAP Beauty Index & MarkPlus. *ZAP Clinic*, 1–36.
- Zhang, J., Chen, C., Liu, P., He, C., & Leung, C. W. K. (2020). Target-guided structured attention network for target-dependent sentiment analysis. *Transactions of the Association for Computational Linguistics*, 8, 172–182. [https://doi.org/10.1162/tacl\\_a\\_00308](https://doi.org/10.1162/tacl_a_00308)