

Analisis Sentimen Berbasis Aspek untuk Mendukung Pengambilan Keputusan pada Industri Perhotelan

DOI: <http://dx.doi.org/10.35889/jutisi.v15i4.3803>

Creative Commons License 4.0 (CC BY – NC)

Kornelius Vincent Christevent^{1*}, Ivan Michael Siregar², Tamsir Hasudungan Sirait³

Sistem Informasi, Institut Teknologi Harapan Bangsa, Bandung, Indonesia

*e-mail Corresponding Author: kornelius.vcent04@gmail.com

Abstract

Guest reviews on Online Travel Agents have become an important source for assessing hotel service quality, yet conventional sentiment analysis only produces an overall score and fails to capture opinions on specific service aspects, making it difficult to determine improvement priorities. This study aims to build an Aspect-Based Sentiment Analysis (ABSA) framework that turns reviews into actionable service-improvement recommendations. Reviews from several platforms were collected, labeled, and classified using a fine-tuned IndoBERT model to identify seven service aspects along with their sentiment polarity, after which the results were summarized and presented through an interactive monitoring dashboard. The proposed model achieved 92.83% accuracy, 92.89% weighted F1, and 74.55% macro F1. These results show that the proposed framework delivers more interpretable and targeted insights than conventional sentiment analysis, thereby supporting data-driven decision-making for service improvement in the hospitality industry.

Keywords: Aspect-Based Sentiment Analysis; IndoBERT; Hotel Reviews; Monitoring Dashboard; Online Travel Agent

Abstrak

Ulasan tamu di *Online Travel Agent* menjadi sumber penting untuk menilai kualitas layanan hotel, namun analisis sentimen konvensional hanya menghasilkan skor keseluruhan dan gagal menangkap opini pada aspek layanan tertentu, sehingga sulit menentukan prioritas perbaikan. Penelitian ini bertujuan membangun kerangka kerja Analisis Sentimen Berbasis Aspek (ABSA) yang mengubah ulasan menjadi rekomendasi perbaikan layanan yang dapat ditindaklanjuti. Ulasan dari beberapa platform dikumpulkan, dilabeli, lalu diklasifikasikan menggunakan model IndoBERT yang disempurnakan untuk mengenali tujuh aspek layanan beserta polaritas sentimennya, kemudian hasilnya diringkas dan disajikan melalui *dashboard* pemantauan interaktif. Model yang diusulkan mencapai akurasi 92,83%, F1 tertimbang 92,89%, dan F1 makro 74,55%. Hasil ini menunjukkan bahwa kerangka kerja yang diusulkan memberikan wawasan yang lebih interpretatif dan tepat sasaran dibanding analisis sentimen konvensional, sehingga mendukung pengambilan keputusan perbaikan layanan berbasis data di industri perhotelan.

Kata Kunci: Analisis Sentimen Berbasis Aspek; IndoBERT; Ulasan Hotel; Dasbor Pemantauan; Agen Perjalanan Daring

1. Pendahuluan

Industri pariwisata dan perhotelan merupakan salah satu sektor penyumbang terbesar perekonomian Indonesia, dengan kontribusi sekitar 4,67% terhadap Produk Domestik Bruto [1] dan penyerapan 25,01 juta tenaga kerja nasional [2]. Jawa Barat menjadi salah satu provinsi dengan kunjungan wisatawan tertinggi, mencapai 88 juta wisatawan pada tahun 2024 dan merupakan angka tertinggi sejak 2019 [3]. Pertumbuhan ini membuat persaingan antarhotel semakin ketat, sehingga menjaga reputasi melalui kualitas layanan operasional menjadi kebutuhan strategis. Ulasan tamu yang dipublikasikan pada *platform Online Travel Agent* (OTA) menjadi cermin langsung pengalaman menginap sekaligus faktor penting dalam keputusan

pemesanan, dengan 87,5% wisatawan Indonesia membaca ulasan sebelum memesan dan 84,4% di antaranya menyatakan ulasan OTA sangat memengaruhi keputusan mereka [4].

Objek penelitian ini adalah satu jaringan hotel yang mengelola enam cabang di Jawa Barat, yaitu Bandung, Bekasi, Bogor, Cirebon, Depok, dan Tasikmalaya, yang menerima ribuan ulasan dari berbagai platform OTA setiap bulan. Volume ulasan yang besar dan tersebar di banyak sumber menyulitkan pemantauan yang konsisten apabila dilakukan tanpa bantuan sistem otomatis. Informasi yang hanya berupa rating bintang juga tidak memadai sebagai dasar pengambilan keputusan operasional karena rating tidak selalu berkorelasi dengan sentimen yang terkandung dalam teks ulasan [5]. Selain itu, analisis sentimen konvensional yang menilai ulasan secara keseluruhan hanya menghasilkan informasi pada tingkat dokumen dan tidak mampu mengungkapkan opini pelanggan terhadap aspek layanan tertentu, sehingga wawasan yang dihasilkan kurang mendalam dan sulit digunakan untuk menentukan prioritas perbaikan secara spesifik.

Untuk mengatasi keterbatasan tersebut, sejumlah penelitian telah mengembangkan Analisis Sentimen Berbasis Aspek atau *Aspect-Based Sentiment Analysis* (ABSA) yang mengaitkan setiap sentimen dengan aspek tertentu yang dibahas dalam ulasan. Secara spesifik, Benlahbib dan Nfaoui [6] serta Mehraliyev et al. [7] meneliti ulasan perhotelan menggunakan klasifikasi tingkat aspek, dan membuktikan bahwa memecah ulasan ke dalam aspek seperti pelayanan, kebersihan, dan fasilitas memberikan informasi yang jauh lebih kaya dibanding skor rating agregat. Dalam konteks ulasan berbahasa Indonesia, Yulianti dan Nissa [8] mengusulkan metode *deep learning* berbasis IndoBERT untuk ABSA dan melaporkan kinerja 23,6% lebih baik dibandingkan Word2Vec, CNN, dan XGBoost. Metode serupa juga dieksplorasi oleh Apriliani et al. [9] yang mencapai akurasi 95,28% pada ulasan hotel bintang lima, serta Perwira et al. [10] yang menggunakan *fine tuning* IndoBERT dan menghasilkan presisi deteksi aspek sebesar 89% pada ulasan pariwisata. Meskipun metode-metode tersebut menunjukkan performa yang tinggi, masih terdapat *gap* utama, sebagian besar penelitian berikut [8], [9], [10] hanya berhenti pada tahap pengembangan dan evaluasi model saja. Riset-riset tersebut belum mentransformasikan hasil sentimen menjadi rekomendasi operasional yang dapat langsung ditindaklanjuti oleh manajemen, sehingga menyisakan kesenjangan antara *output* metrik model dan kebutuhan nyata pengambilan keputusan di lapangan.

Berdasarkan kesenjangan tersebut, penelitian ini mengusulkan kerangka kerja ABSA berbasis IndoBERT *multi-head* yang tidak hanya mengklasifikasikan sentimen pada tingkat aspek, tetapi juga mengubahnya menjadi wawasan rekomendasi perbaikan layanan dan menyajikannya melalui *dashboard* pemantauan interaktif. Pemilihan IndoBERT didukung oleh riset yang membuktikan keunggulan mekanisme *self-attention* dua arahnya dalam memahami konteks semantik dan menangani bahasa campuran serta istilah informal yang sangat lazim pada ulasan tamu berbahasa Indonesia [11], [12]. Selain itu, adopsi konsep *multi-head* pada model ini secara teori memungkinkan klasifikasi multiaspek dilakukan secara serentak, yang terbukti jauh lebih efisien [13]. Selanjutnya, penyajian melalui *dashboard* diusulkan sebagai bagian krusial dari solusi berdasarkan kajian bahwa hasil algoritma data berskala besar hanya akan bernilai strategis apabila divisualisasikan dalam antarmuka yang intuitif [14], [15]. Hal ini secara langsung memperkuat temuan Wuisang dan Pangestu [16] yang menegaskan bahwa hasil analisis sentimen baru berdampak nyata ketika diintegrasikan ke dalam antarmuka operasional yang siap digunakan. Dengan demikian, kebaruan penelitian ini berfokus pada penyatuan keandalan model ABSA IndoBERT dengan *dashboard monitoring* ke dalam satu kerangka terintegrasi, yang secara efektif menjawab masalah utama studi sebelumnya dengan menjembatani jarak antara evaluasi akurasi algoritma dan eksekusi perbaikan manajemen.

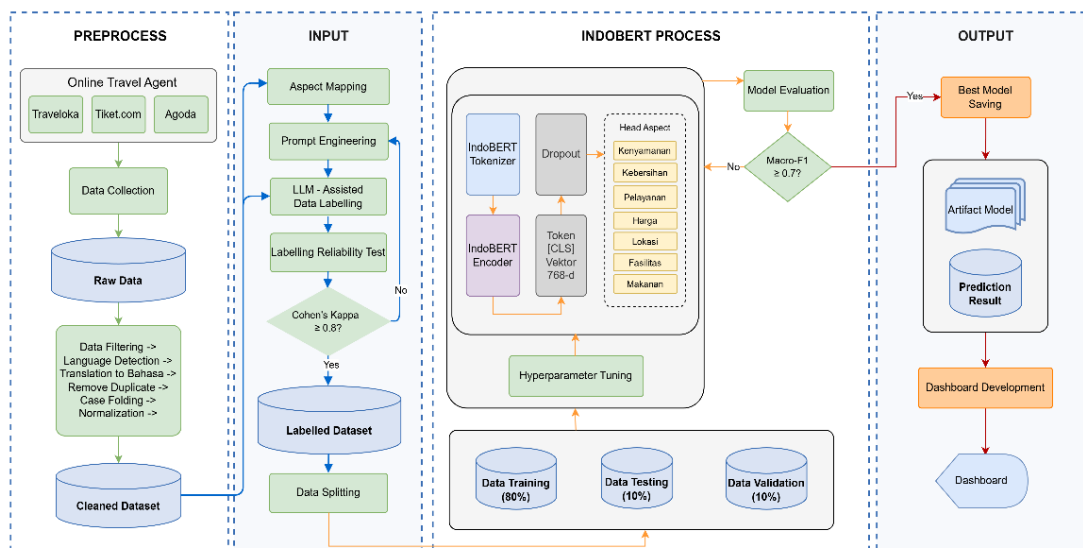
Dataset penelitian ini diperoleh dari ulasan pelanggan pada enam cabang hotel tersebut yang dikumpulkan dari tiga platform OTA, yaitu Traveloka, Agoda, dan Tiket.com. Adapun kontribusi utama penelitian ini adalah sebagai berikut:

- 1) Mengembangkan dan memvalidasi *framework* ABSA berbasis IndoBERT untuk mengidentifikasi aspek layanan hotel dan menentukan polaritas sentimen pada setiap aspek menggunakan dataset ulasan berbahasa Indonesia yang dikumpulkan dari tiga platform OTA.
- 2) Merancang mekanisme pembentukan rekomendasi berbasis agregasi sentimen aspek untuk mengidentifikasi kekuatan dan kelemahan layanan, serta menentukan prioritas perbaikan operasional berdasarkan tingkat kepentingan dan persepsi pelanggan.

- 3) Mengimplementasikan *framework* yang diusulkan ke dalam *dashboard* interaktif yang menyajikan hasil analisis dan rekomendasi secara visual, sehingga mendukung *monitoring* kualitas layanan dan pengambilan keputusan berbasis data bagi manajemen hotel.

2. Metodologi

Penelitian ini memadukan kerangka kerja CRISP-ML(Q) [17] untuk pembangunan model dan pengembangan *dashboard*. Kerangka kerja penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1 Kerangka kerja ABSA-IndoBERT Rekomendasi Perbaikan Layanan Hotel

2.1 Pre-Process

a. Data Collection

Pada fase ini, data ulasan tamu hotel dikumpulkan dari tiga platform OTA yaitu Traveloka, Agoda, dan Tiket terhadap sebuah hotel yang berlokasi di beberapa daerah di Jawa Barat yaitu Bandung, Bekasi, Bogor, Cirebon, Depok, dan Tasikmalaya [4]. Pengambilan data dilakukan dengan JavaScript melalui browser *DevTools Console* dan disimpan dalam format CSV. Hasilnya dari masing-masing OTA dan lokasi selanjutnya dilakukan agregasi ke dalam dataset tunggal dan standarisasi ke dalam format yang memiliki kolom ID_Ulasan, Platform, Nama_Hotel, Review_Date, Teks_Review seperti yang terdapat pada Tabel 1.

Tabel 1 Contoh Data Ulasan Mentah dan Hasil *Preprocessing*

(a) ID_Ulasan	(b) Platform	(c) Nama_Hotel	(d) Tanggal_Review	(e) Teks_Ulasan (Data Mentah)	(f) Teks_Ulasan (Data Bersih)
R-001	Agoda	Hotel Bandung	1/8/2025	Staff super ramah dan lokasi yg sangat dekat kemana2 sangat cocok untuk membawa keluarga	staff super ramah dan lokasi yang sangat dekat kemana-mana sangat cocok untuk membawa keluarga
R-004	Tiket.com	Hotel Bogor	11/26/2021	cita rasa makanannya ditingkatkan 🍕	cita rasa makanannya ditingkatkan
R-005	Agoda	Hotel Bandung	10/20/2017	Very highly recommended, staff was nice, room is spacious, clean, and comfy. Breakfast not as good as all reviewers said...	sangat direkomendasikan, staf ramah, kamar luas, bersih, dan nyaman. sarapan tidak sebaik yang dikatakan pengulas lain...

b. Data Preprocessing

Selanjutnya dilakukan pembersihan terhadap dataset melalui enam tahap berurutan konteks kalimat tetap utuh untuk IndoBERT [18].

- 1) *Data Filtering*. Ulasan kosong dan berpanjang kurang dari 20 karakter dibuang dengan penyaringan pantas, karena terlalu pendek untuk memuat opini bermakna.
- 2) *Language Detection*. Bahasa tiap ulasan dideteksi memakai pustaka *lingua-language-detector* (19 bahasa, ambang keyakinan 0,50, teks terlalu pendek diberi label *unknown*). Tahap ini memberi label bahasa, bukan membuang data.
- 3) *Translation to Bahasa*. Ulasan non-Indonesia diterjemahkan ke Bahasa Indonesia memakai *deep-translator* (Google Translate), dengan teks asli disimpan dan dikembalikan bila salah deteksi, agar seragam untuk IndoBERT.
- 4) *Remove Duplicate*. Ulasan yang sama persis dibuang dengan penghapusan duplikat pantas, dilakukan sebelum dan sesudah normalisasi agar satu opini tidak terhitung ganda.
- 5) *Case Folding*. Seluruh huruf diubah menjadi huruf kecil agar kata yang sama tidak dianggap berbeda.
- 6) *Normalization*. Memakai ekspresi reguler (modul *re*) dan kamus slang untuk membuang emoji, karakter kontrol, serta simbol berlebih (~ * # @ ^ & | \ { } [] < >), menyeragamkan tanda kutip dan spasi, membakukan singkatan dan kata gaul, dan memangkas karakter berulang [19].

Langkah *stopword* dan *stemming* tidak dilakukan karena IndoBERT mengandalkan konteks kalimat utuh, dan tokenisasi akhir ditangani *tokenizer* bawaan IndoBERT.

2.2 Input

Tahap Input bertujuan menghasilkan kumpulan ulasan berlabel yang dapat dipercaya sebagai bahan pelatihan model. Prosesnya dijelaskan sebagai berikut.

a. Aspect Mapping

Proses ini menentukan tujuh aspek yang dinilai dari ulasan hotel, yaitu Lokasi, Kenyamanan, Pelayanan, Kebersihan, Harga, Makanan, dan Fasilitas. Setiap aspek diberi penjelasan dan contoh kata penanda, serta aturan untuk membedakan aspek yang mirip. Sebagai contoh, komentar tentang kondisi kamar dimasukkan ke Kenyamanan, sedangkan keberadaan atau kerusakan perangkat dimasukkan ke Fasilitas. Pembagian aspek ini disusun dari kajian penelitian terdahulu tentang ulasan hotel agar sesuai dengan hal yang benar-benar diperhatikan tamu [7], [20].

b. Prompt Engineering

Proses ini menyusun perintah yang memandu *Large Language Model (LLM)* saat memberi label. Perintah berisi penjelasan tugas dan aturan yang jelas, seperti hanya boleh memakai tujuh aspek dan tiga jenis sentimen yaitu positif, negatif, dan netral, hanya menilai aspek yang benar-benar disebut dalam ulasan, serta menuliskan hasil dalam bentuk yang rapi dan seragam. Perintah juga mengatur cara menanganai kalimat yang mengandung penyangkalan dan ulasan yang menyebut beberapa aspek sekaligus. Perintah yang dirancang dengan baik membuat hasil pelabelan menjadi lebih konsisten [21].

c. LLM-Assisted Data Labelling

Pemberian label dilakukan secara otomatis oleh model *LLM* yaitu *Claude Sonnet 4.6* dengan mengikuti perintah yang telah dibuat. Model membaca setiap ulasan, lalu menentukan sentimen pada tiap aspek beserta alasan singkatnya. Cara ini dipilih karena jumlah ulasan sangat banyak sehingga pelabelan manual akan memakan waktu yang lama, sementara model mampu bekerja jauh lebih cepat dengan hasil yang seragam. *Output* nya berupa kumpulan ulasan berlabel yang setiap barisnya memuat informasi ulasan, sentimen untuk ketujuh aspek, dan alasannya, sehingga langsung siap dipakai untuk melatih model [22], [23].

d. Labelling Reliability Test (Cohen's Kappa)

Proses ini menguji apakah label dari model dapat dipercaya, dengan membandingkannya terhadap label yang dibuat seorang ahli pada sebagian ulasan. Tingkat kesesuaian keduanya diukur menggunakan *Cohen's Kappa* [24].

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

dengan P_o adalah tingkat kesesuaian yang teramati dan P_e adalah tingkat kesesuaian yang mungkin terjadi secara kebetulan. Label dianggap dapat dipercaya bila nilai Kappa mencapai minimal 0,8, yang menurut skala Landis dan Koch tergolong tingkat kesesuaian hampir sempurna. Bila nilai belum mencapai batas tersebut, perintah pada tahap Prompt Engineering diperbaiki dan pelabelan diulang. Bila sudah mencapai batas, proses dilanjutkan ke pembagian data.

e. Data Splitting

Proses ini membagi kumpulan ulasan berlabel menjadi data latih, data validasi, dan data uji dengan proporsi 80%, 10%, dan 10%. Data latih dipakai untuk melatih model, data validasi untuk memantau kinerja dan menyetel pengaturan selama pelatihan, dan data uji untuk mengukur kinerja akhir pada ulasan yang belum pernah dilihat model [17].

2.3 IndoBERT Process

Tahap ini merupakan inti pemodelan, yaitu mengubah ulasan menjadi prediksi sentimen pada tiap aspek menggunakan model IndoBERT yang di *fine-tuning* pada data berlabel.

a. Tokenization

Sebelum masuk ke model, teks ulasan dipecah menjadi potongan kecil yang disebut *token* oleh *tokenizer* bawaan IndoBERT dengan metode *WordPiece*. Kata yang jarang muncul dipecah menjadi *sub-word* sehingga model tetap dapat mengenalinya, dan setiap ulasan disamakan panjangnya hingga batas *max length* tertentu agar dapat diproses bersama dalam satu *batch* [11].

b. IndoBERT Encoder

Token kemudian diolah oleh *encoder* IndoBERT, yaitu model bahasa yang sudah dilatih pada teks Bahasa Indonesia dalam jumlah besar. *Encoder* membaca seluruh kata dalam kalimat secara bersamaan sehingga memahami makna kata berdasarkan konteksnya. Hasilnya berupa *embedding*, yaitu representasi angka yang merangkum isi ulasan dan diwakili oleh sebuah *token* khusus di awal kalimat [11], [23].

c. Dropout

Sebelum tahap klasifikasi, sebagian nilai pada *embedding* tersebut dinonaktifkan sementara secara acak menggunakan *dropout*. Langkah ini mencegah *overfitting*, yaitu kondisi model terlalu menghafal *training data*, sehingga model tetap andal ketika menghadapi ulasan baru [21].

d. Aspect Heads

Model memakai satu *encoder* bersama yang terhubung ke tujuh *classifier head*, satu untuk setiap aspek. Setiap *head* menghasilkan satu dari empat kemungkinan *output*, yaitu tidak disebut, positif, negatif, atau netral. Dengan rancangan ini, satu ulasan dapat dinilai pada ketujuh aspek sekaligus dalam sekali proses [13].

e. Hyperparameter Tuning

Untuk memperoleh kombinasi pengaturan terbaik, dilakukan percobaan terhadap beberapa nilai *hyperparameter* seperti varian IndoBERT, *learning rate*, *batch size*, jumlah *epoch*, dan *max length*. Pelatihan memakai *optimizer* AdamW dengan *learning rate scheduler*, ditambah *class weighting* untuk mengatasi data yang timpang dan *early stopping* agar pelatihan berhenti saat kinerja tidak lagi membaik. Kombinasi dengan kinerja validasi terbaik dipilih sebagai model akhir [25].

f. Model Evaluation

Kinerja model akhir diukur pada *test data* menggunakan *accuracy*, *macro F1-score*, dan *weighted F1-score*. *Accuracy* menunjukkan proporsi prediksi yang benar, *macro F1-score* menilai

rata-rata kinerja semua kelas secara setara sehingga adil bagi kelas minoritas, dan *weighted F1-score* menimbang kinerja sesuai jumlah data pada tiap kelas. Ketiganya dipakai bersama karena distribusi sentimen pada data tidak seimbang [8].

2.4 Output

Tahap Output mengambil model terbaik dari tahap sebelumnya, menyimpannya, menggunakannya untuk memprediksi seluruh ulasan, lalu menyajikan hasilnya melalui *dashboard* interaktif sebagai pendukung pengambilan keputusan. Rangkaian prosesnya dijelaskan berikut.

a. Best Model Saving

Setelah proses *hyperparameter tuning* dan evaluasi selesai, model dengan kinerja validasi terbaik dipilih dan disimpan sebagai model final. Model disimpan dalam bentuk *artifact*, yang terdiri dari bobot model, *tokenizer*, dan *metadata*, sehingga dapat dimuat ulang kapan pun tanpa perlu dilatih dari awal. *Artifact* ini kemudian dipakai untuk memprediksi seluruh ulasan pada dataset sekaligus melalui *batch inference*, dan hasilnya berupa label sentimen tiap aspek yang disimpan sebagai *prediction result*, yang menjadi sumber data utama bagi *dashboard* [11], [17].

b. Dashboard Development

Pada proses ini dibangun *dashboard* interaktif menggunakan Flask dan grafik Chart.js berdasarkan *prediction result*. Hasil prediksi diringkas melalui proses *aggregation* dan aspek dengan sentimen negatif yang konsisten diberi prioritas, sehingga data sentimen dapat diterjemahkan menjadi rekomendasi perbaikan layanan yang terarah. *Dashboard* yang dihasilkan menyajikan ringkasan sentimen per aspek, perbandingan antar cabang dan platform, serta tren waktu, dan juga menyediakan halaman prediksi sentimen secara *real-time* untuk menilai ulasan baru. Penyajian visual ini memudahkan manajemen membaca temuan dan mengambil keputusan berbasis data [6], [14], [15], [26].

3. Hasil dan Pembahasan

3.1 Preprocess

Ulasan dikumpulkan dari tiga platform OTA, yaitu Traveloka, Agoda, dan Tiket.com, untuk enam cabang Hotel ini di Jawa Barat (Bandung, Bekasi, Bogor, Cirebon, Depok, dan Tasikmalaya), dan menghasilkan 18.297 ulasan mentah. Berdasarkan platform, Traveloka menyumbang ulasan terbanyak (8.863), diikuti Agoda (6.670) dan Tiket.com (2.764). Setelah melewati enam tahap pra-pemrosesan, jumlahnya berkurang menjadi 15.095 ulasan bersih, sehingga 3.202 ulasan (17,5%) dibuang karena duplikat, terlalu pendek, atau menjadi kosong setelah dibersihkan. Pengurangan terbesar terjadi pada ulasan Agoda, yang turun menjadi 3.895 karena paling banyak mengandung duplikat dan ulasan singkat tanpa opini bermakna, sedangkan Traveloka dan Tiket.com relatif stabil di angka 8.825 dan 2.375.

Hasil akhir tiap ulasan dapat dilihat pada kolom (f) Tabel 1. Kolom (a) sampai (e) pada tabel tersebut merupakan data mentah yang langsung diperoleh saat pengumpulan, sedangkan kolom (f) belum tersedia pada tahap itu dan baru dihasilkan pada tahap ini, yaitu setelah teks pada kolom (e) melewati keenam tahap pra-pemrosesan. Dengan demikian kolom (f) memperlihatkan bentuk akhir ulasan yang siap dilabeli, sebagaimana dicontohkan pada baris R-001, R-004, dan R-005.

3.2 Input

3.2.1 Aspect Mapping

Tujuh aspek ditetapkan melalui triangulasi tiga sumber, yaitu dukungan literatur, kata kunci dominan, dan pertimbangan *expert*, agar definisinya konsisten dengan konteks Hotel ini. Pemetaan ini disandingkan dengan hasil distribusi label per aspek pada Tabel 3.

Tabel 2 Pemetaan taksonomi dan distribusi label per aspek

Aspek	Taksonomi Aspek			Polaritas			Total	Kappa
	Dukungan Literatur	Kata Kunci Dominan	Pertimbangan Expert	Positif	Negatif	Netral		
Pelayanan	Keramahan staf dan loyalitas [5], [16].	staf, ramah, pelayanan	Indikator kinerja operasional	5.070	875	362	6.307	0,8703
Kenyamanan	Kondisi kamar penentu kepuasan [7], [16].	kamar, nyaman, luas	Tolok ukur kualitas kamar	3.985	1.619	259	5.863	0,8706
Makanan	Kualitas sarapan sebagai pembeda [7], [20].	sarapan, makanan, enak	Nilai jual penting Hotel	3.875	915	566	5.356	0,9018
Lokasi	Posisi strategis dan aksesibilitas [7], [16].	lokasi, strategis, dekat	Dekat tempat fungsional tertentu	4.192	282	192	4.666	0,8474
Kebersihan	Higiene krusial pascapandemi [5], [26].	bersih, kebersihan, rapi	Standar mutu seluruh cabang	3.556	616	73	4.245	0,8462
Fasilitas	Kelengkapan penunjang kenyamanan [16], [20].	kolam, parkir, fasilitas	Sarana tertentu yang disediakan	1.465	1.589	561	3.615	0,8320
Harga	Kesepadanan nilai dan tarif [5], [26].	harga, mahal, terjangkau	Dasar strategi penetapan tarif	639	254	138	1.031	0,8337
Total				22.782	6.150	2.151	31.083	0,8574

3.2.2 Prompt Engineering

Instruksi pelabelan dirancang agar model bekerja seperti *annotator* ahli. *Prompt* memuat definisi ketujuh aspek beserta kata sinyal dan aturan pembeda untuk kasus mirip, misalnya "AC dingin" tergolong Kenyamanan sedangkan "AC rusak" tergolong Fasilitas. *Prompt* juga menetapkan tiga polaritas, aturan hanya melabeli aspek yang disebut eksplisit, maksimal satu entri per aspek, serta penanganan negasi, opini campuran, dan kalimat multi-aspek. Beberapa contoh *few-shot* disertakan sebagai acuan [21], dan *output* diatur dalam format tabel 19 kolom berisi review, label tiap aspek, dan alasannya.

3.2.3 LLM-Assisted Data Labelling

Pelabelan menggunakan *Large Language Model*, dan beberapa kandidat dibandingkan terlebih dahulu, yaitu Claude Sonnet 4.6, GPT-4o, Gemini 2.5 Pro, dan Llama 3.3 70B. Claude Sonnet 4.6 dipilih karena memiliki cakupan label tertinggi 74,86% dan kepatuhan instruksi terbaik, dua kriteria yang paling menentukan pada anotasi ABSA multi-aspek. Model lain seperti GPT-4o memang mencatat akurasi mentah lebih tinggi, tetapi cakupan labelnya lebih rendah sehingga banyak aspek terlewat, sedangkan Gemini 2.5 Pro jauh lebih lambat untuk skala besar [22], [23]. Dengan model tersebut, 15.095 ulasan dilabeli dan menghasilkan 31.083 label aspek (positif 73,3%, negatif 19,8%, netral 6,9%). Kualitas label diuji dengan *Cohen's Kappa* pada 150 ulasan dan memperoleh rata-rata makro 0,8574, tergolong kuat menurut Landis dan Koch, sehingga hasil pelabelan dinilai baik. Aspek Fasilitas memiliki proporsi sentimen negatif tertinggi sehingga menjadi perhatian khusus pada pembahasan berikutnya [24].

Sebagai ilustrasi, salah satu ulasan diambil yang berisi: "*kamar sesuai dengan foto, luas, cukup bersih, sarapan enak dan lokasi strategis, dekat dengan tempat-tempat makan. hanya saja wifi kurang bagus*" Dari satu ulasan ini Claude Sonnet 4.6 memberi label sentimen yang berbeda pada masing-masing aspek beserta alasannya, seperti pada Tabel 3.

Tabel 3 Contoh hasil pelabelan satu ulasan dengan sentimen campuran

Aspek	Sentimen	Alasan
Kenyamanan	Positif	Kamar sesuai dengan foto, luas
Kebersihan	Positif	Cukup bersih
Lokasi	Positif	Lokasi strategis, dekat dengan tempat-tempat makan
Fasilitas	Negatif	Wifi kurang bagus
Makanan	Positif	Sarapan enak

3.2.4 Data Splitting

Data berlabel dibagi menjadi data latih, validasi, dan uji dengan perbandingan 80/10/10 menggunakan teknik *multi-label stratified split* agar proporsi label tiap aspek tetap seimbang di setiap subset. Pembagian ini menghasilkan 12.075 ulasan untuk pelatihan, 1.510 ulasan untuk validasi, dan 1.510 ulasan untuk pengujian, dari total 15.095 ulasan berlabel [17].

3.3 IndoBERT Process

Melalui *hyperparameter tuning*, model terbaik diperoleh menggunakan IndoBERT varian *base-p1* dengan panjang token maksimum 192, ukuran *batch* 8 (akumulasi gradien 2), *learning rate* 2×10^{-5} , dan 10 *epoch*. Pelatihan memakai *optimizer* AdamW dengan *weight decay* 0,01, *learning rate scheduler* berbasis *cosine* dan *warmup* 10%, *dropout* 0,2, *label smoothing* 0,05, serta *class weighting* aktif untuk mengatasi distribusi label yang timpang [11], [25]. Pada data uji, model mencapai akurasi 92,83%, *weighted F1-score* 92,89%, dan *macro F1-score* 74,55%. Nilai akurasi dan *weighted F1* yang tinggi menunjukkan model sangat baik mengenali kelas mayoritas, sedangkan *macro F1* yang lebih rendah mencerminkan tantangan pada kelas minoritas, terutama sentimen netral yang jumlahnya sedikit. Kinerja model untuk setiap aspek ditunjukkan pada Tabel 4 [8].

Tabel 4 Kinerja model per aspek (%)

Aspek	Macro F1	Weighted F1	Accuracy
Makanan	78,10	93,55	93,32
Kebersihan	77,00	95,71	95,70
Harga	75,78	97,72	97,75
Fasilitas	74,43	89,89	89,61
Lokasi	73,44	95,33	95,50
Pelayanan	72,58	91,14	91,07
Kenyamanan	70,49	86,93	86,90
Rata-rata	74,55	92,89	92,83

Aspek Makanan, Kebersihan, dan Harga memperoleh *macro F1* tertinggi karena pola sentimennya relatif jelas dan konsisten, sedangkan Kenyamanan memiliki *macro F1* terendah karena opini tamu pada aspek ini lebih beragam dan sering bercampur sehingga lebih sulit dibedakan. Secara keseluruhan, kinerja model sudah memadai untuk dipakai pada tahap penyajian hasil melalui *dashboard*.

3.4 Output

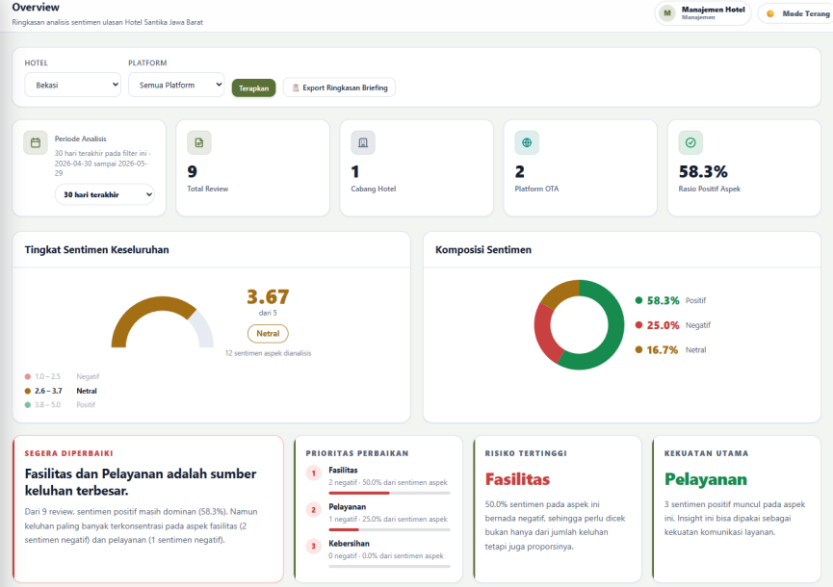
3.4.1 Implementasi Dashboard

Hasil akhir berupa *dashboard* interaktif yang dibangun dengan Flask dan grafik Chart.js, dengan model IndoBERT dimuat saat aplikasi dijalankan dan seluruh ulasan diprediksi secara *batch* sebagai sumber datanya. *Dashboard* memuat halaman *Overview*, analisis per aspek, perbandingan antarcabang dan antarplatform, tren waktu, eksplorasi ulasan, performa model, serta halaman prediksi sentimen *real-time* untuk menilai ulasan baru. Penyajian visual ini memudahkan manajemen membaca temuan dan mengambil keputusan berbasis data [14], [15].

3.4.2 Analisis Insight Dashboard untuk Strategi Perbaikan

Melalui *dashboard*, hasil klasifikasi dapat ditelusuri menurut cabang, platform, aspek, dan periode, sehingga manajemen dapat memfokuskan perbaikan pada aspek yang paling

bermasalah di satu cabang dan rentang waktu tertentu, bukan sekadar membaca skor rata-rata. Sebagai simulasi, *dashboard* difilter untuk Hotel Bekasi pada seluruh platform selama 30 hari terakhir (30 April sampai 29 Mei 2026). Pada periode ini terdapat 9 ulasan dengan 12 sentimen aspek, skor keseluruhan 3,67 dari 5 (tergolong netral), dan komposisi 58,3% positif, 25,0% negatif, serta 16,7% netral (Gambar 2). Meskipun sentimen positif masih dominan, panel prioritas menunjukkan keluhan terkonsentrasi pada aspek Fasilitas (2 sentimen negatif, rasio 50,0%) dan Pelayanan (1 sentimen negatif, rasio 25,0%), sehingga keduanya ditetapkan sebagai prioritas perbaikan (Gambar 3).



Gambar 2 Halaman Overview untuk Hotel Bekasi

Tabel Prioritas Perbaikan						
Cabang dan aspek dengan keluhan negatif tertinggi sebagai bahan briefing layanan.						
Menampilkan 1-2 dari 2 prioritas						
NO.	HOTEL	ASPEK	JUMLAH NEGATIF	RASIO NEGATIF	PLATFORM DOMINAN	REVIEW PENDUKUNG TERBARU
1	Bekasi Hotel Sankia Bekasi	Fasilitas	2 dari 4 sentimen aspek	50.0%	Tiket	5/13/2026 - Tiket receptionist cepat tanggap, ada welcome drink, kamar dan fasilitas lain standar hotel pada umumnya, makanan lumayan enak namun ketika habis lama untuk refillnya, kapasitas parkir di rooftop sedikit, selengkapnya
2	Bekasi Hotel Sankia Bekasi	Pelayanan	1 dari 4 sentimen aspek	25.0%	Traveloka	5/13/2026 - Traveloka overall baik, cuma 1 aja minumannya, saya minta tambahan air mineral, kan katanya free 2 botol lagi per hari dari awal check in sampai besoknya tidak diantar, sudah telepon sampai 2x terus bilang ke staf yang antar room service juga.

Gambar 3 Tabel Prioritas Perbaikan untuk Hotel Bekasi

Pada Fasilitas, keluhan berkaitan dengan ketiadaan lemari es di kamar dan kapasitas parkir *rooftop* yang terbatas. Pada Pelayanan, keluhan bukan menandakan layanan yang buruk karena resepsionis tetap dinilai ramah dan cepat, melainkan komitmen pemberian air mineral gratis yang tidak dipenuhi meskipun tamu sudah meminta beberapa kali. Contoh ulasan pendukung beserta keluhan prioritas dan strategi perbaikannya disajikan secara ringkas pada Tabel 5.

Tabel 5 Ulasan pendukung, keluhan prioritas, dan strategi perbaikan

ID	Platform	Cuplikan Ulasan	Keluhan Prioritas (Aspek)	Strategi Perbaikan
R-4792	Tiket.com	"bagus bersih ramah, minusnya ngga ada lemari es di kamar"	Fasilitas: kamar tidak dilengkapi lemari es	Melengkapi kulkas atau lemari es mini pada kamar
R-4863	Tiket.com	"receptionist cepat tanggap, ada welcome drink, kamar dan fasilitas lain standar hotel pada umumnya, kapasitas parkir di rooftop sedikit"	Fasilitas: kapasitas parkir <i>rooftop</i> terbatas	Mengevaluasi penataan atau menambah kapasitas parkir

ID	Platform	Cuplikan Ulasan	Keluhan Prioritas (Aspek)	Strategi Perbaikan
R-11821	Traveloka	"saya minta tambahan air mineral, katanya gratis 2 botol per hari, tetapi dari awal check-in sampai besoknya tidak diantar walau sudah telepon sampai 3 kali"	Pelayanan: komitmen amenities tidak terpenuhi dan tindak lanjut permintaan tamu lambat	Memperketat prosedur <i>room service</i> dan pemenuhan amenities, serta memantau waktu respons

3.4.3 Pembahasan

Hasil pengujian menunjukkan model ABSA berbasis IndoBERT *multi-head* mencapai akurasi 92,83%, *weighted F1-score* 92,89%, dan *macro F1-score* 74,55% pada data uji. Capaian ini menjawab tujuan pertama penelitian, yaitu membangun kerangka kerja yang mampu mengidentifikasi tujuh aspek layanan sekaligus menentukan polaritas sentimennya dalam satu kali proses inferensi. Nilai akurasi dan *weighted F1* yang tinggi menandakan model sangat baik mengenali kelas mayoritas, sedangkan *macro F1* yang lebih rendah mencerminkan tantangan pada kelas minoritas. Perbedaan kedua metrik ini konsisten dengan sifat *macro F1* yang memberi bobot setara pada setiap kelas, sehingga lebih sensitif terhadap distribusi data yang timpang dan menjadi indikator yang lebih jujur untuk data ulasan yang secara alami didominasi sentimen tertentu.

Perbedaan kinerja pada semua aspek dapat dijelaskan oleh karakteristik opini pada masing-masing aspek. Aspek Makanan, Kebersihan, dan Harga memperoleh *macro F1* tertinggi karena kosakata dan pola sentimennya relatif eksplisit dan konsisten, sedangkan aspek Kenyamanan memperoleh nilai terendah karena opini tamu pada aspek ini cenderung beragam dan sering bercampur dalam satu kalimat sehingga lebih sulit dibedakan. Rendahnya pengenalan kelas netral juga disebabkan jumlahnya yang minoritas, dan penerapan *class weighting* serta label smoothing terbukti membantu model tetap memberi perhatian pada kelas yang jarang muncul. Temuan ini sejalan dengan Mehraliyev et al. [7] yang menunjukkan bahwa dekomposisi ulasan ke tingkat aspek menghasilkan informasi yang lebih granular dibanding penilaian menyeluruh. Dibandingkan Miranda [27] yang menggunakan pendekatan *machine learning* klasik untuk ABSA, kerangka kerja berbasis IndoBERT pada penelitian ini mampu menangkap konteks dua arah sehingga lebih tangguh pada ulasan dengan sentimen campuran, sekaligus mengukuhkan pemilihan IndoBERT sebagai *backbone* model [12].

Simulasi penelusuran untuk Hotel Bekasi memperlihatkan bahwa agregasi sentimen tingkat aspek mampu menyingkap prioritas perbaikan yang tidak terlihat dari skor rata-rata. Meskipun skor keseluruhan tergolong positif (3,67 dari 5 dengan 58,3% sentimen positif), tabel prioritas tetap mengidentifikasi aspek Fasilitas dan Pelayanan sebagai titik keluhan yang terkonsentrasi. Hal ini menjawab tujuan kedua dan ketiga penelitian, yaitu mengubah *output* model menjadi rekomendasi yang dapat ditindaklanjuti dan menyajikannya melalui *dashboard* interaktif. Temuan ini memperlihatkan kontribusi yang berbeda dari sebagian besar studi terdahulu yang berhenti pada pelaporan akurasi model, karena rasio sentimen negatif per aspek terbukti menjadi sinyal keputusan yang lebih tajam dibanding rating agregat dalam menentukan sasaran perbaikan operasional.

Kualitas *framework* ini juga didukung oleh proses pelabelan. Rata-rata Cohen's Kappa sebesar 0,8574 antara label otomatis dan anotasi pakar menunjukkan tingkat kesepakatan yang kuat, sehingga data latih yang digunakan dapat dipertanggungjawabkan. Hasil ini memperkuat temuan bahwa *Large Language Model* dapat berperan sebagai annotator yang baik untuk tugas berskala besar, sebagaimana ditunjukkan Lindahl [28] pada pemanfaatan LLM untuk teks domain khusus. Efektivitas pelabelan otomatis ini juga tidak lepas dari perancangan instruksi yang cermat, sejalan dengan Liu et al. [29] yang menegaskan bahwa kualitas *prompt* sangat menentukan kualitas *output LLM*, sehingga pendekatan ini menjadi kontribusi metodologis dalam menyediakan data ABSA berbahasa Indonesia secara efisien.

Penelitian ini berhasil mengimplementasikan model ABSA berbasis IndoBERT dan *dashboard* pemantauan ke dalam satu kerangka terintegrasi yang menjembatani jarak antara pengembangan model dan pemanfaatannya di industri perhotelan. Meskipun demikian, penelitian masih memiliki keterbatasan, terutama pada pengenalan kelas netral yang minoritas, cakupan data yang terbatas pada enam cabang dan tiga platform OTA, serta validasi pelabelan yang melibatkan satu orang pakar. Keterbatasan ini membuka peluang penelitian lanjutan, antara

lain augmentasi data dan penyeimbangan distribusi kelas untuk meningkatkan macro F1, pelibatan lebih dari satu anotator pakar untuk memperkuat validitas label, serta pengembangan deteksi aspek implisit agar sistem mampu menangkap opini yang tidak dinyatakan secara eksplisit. Pengembangan ke arah tersebut diharapkan membuat kerangka kerja semakin akurat dan bermanfaat sebagai alat bantu pengambilan keputusan di industri perhotelan.

4. Simpulan

Penelitian ini berhasil membangun sistem analisis sentimen berbasis aspek (ABSA) untuk ulasan Hotel di Jawa Barat pada tujuh aspek layanan, yaitu lokasi, kenyamanan, pelayanan, kebersihan, harga, makanan, dan fasilitas. Pendekatan ini menjawab kebutuhan industri perhotelan untuk memahami opini tamu secara lebih rinci dibandingkan sekadar rating, sehingga penilaian terhadap setiap aspek layanan dapat dipetakan secara terukur.

Penerapan IndoBERT dengan mekanisme klasifikasi multi-aspek mencapai tingkat akurasi dan performa yang sangat baik, yang juga didukung oleh dataset berlabel dengan tingkat kesesuaian dengan pakar yang kuat. Hasil klasifikasi tersebut disajikan melalui dasbor interaktif yang dapat ditelusuri menurut cabang, platform, aspek, dan periode. Sebagai langkah praktis pemanfaatannya, pihak manajemen dapat menggunakan dasbor ini secara berkala untuk mengidentifikasi titik lemah layanan dengan cepat, menyusun materi pengarahan (*briefing*) harian bagi staf, serta mengalokasikan anggaran perbaikan secara spesifik pada fasilitas atau layanan yang menjadi prioritas keluhan tamu. Hal ini mengubah *output* model menjadi wawasan yang langsung dapat ditindaklanjuti secara tepat sasaran.

Penelitian masih memiliki keterbatasan, terutama pada pengenalan kelas netral yang minoritas dan cakupan data yang terbatas pada enam cabang. Penelitian selanjutnya disarankan memperluas cakupan data antarcabang dan antarplatform, menyeimbangkan distribusi kelas, serta mengembangkan analisis sentimen tingkat lanjut seperti deteksi aspek implisit agar sistem semakin akurat dan bermanfaat bagi pengambilan keputusan.

Daftar Referensi

- [1] Kementerian Pariwisata, "Perkembangan PDB Sektor Pariwisata Tahun 2024."
- [2] Kementerian Koordinator Bidang Perekonomian Republik Indonesia, "Mendongkrak Kinerja Sektor Pariwisata sebagai Penggerak Ekonomi Nasional."
- [3] Dinas Pariwisata dan Kebudayaan Provinsi Jawa Barat, "Rilis Data Statistik Pariwisata Jawa Barat Tahun 2024," Bandung, 2024.
- [4] K. Ariansyah, J. Prawiro, and R. Sanjaya, "Pengaruh Ulasan Online Terhadap Keputusan Wisatawan dalam Memilih Hotel," *J. Pariwisata dan Perhotelan*, vol. 2, no. 2, p. 8, Jan. 2025, doi: 10.47134/pjpp.v2i2.3559.
- [5] B. A. Sparks and V. Browning, "The impact of online reviews on hotel booking intentions and perception of trust," *Tour. Manag.*, vol. 32, no. 6, pp. 1310–1323, Dec. 2011, doi: 10.1016/j.tourman.2010.12.011.
- [6] A. Benlahbib and E. H. Nfaoui, "A hybrid approach for generating reputation based on opinions fusion and sentiment analysis," *J. Organ. Comput. Electron. Commer.*, vol. 30, no. 1, pp. 9–27, Jan. 2020, doi: 10.1080/10919392.2019.1654350.
- [7] F. Mehrliyyev, I. C. C. Chan, and A. P. Kirilenko, "Sentiment analysis in hospitality and tourism: a thematic and methodological review," *Int. J. Contemp. Hosp. Manag.*, vol. 34, no. 1, pp. 46–77, Jan. 2022, doi: 10.1108/IJCHM-02-2021-0132.
- [8] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.
- [9] S. Apriliani, A. Erfina, and C. Warman, "Fine-Tuned IndoBERT for Aspect-Based Sentiment Analysis of Indonesian Five-Star Hotel Reviews," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 14, no. 4, pp. 437–445, Oct. 2025, doi: 10.32736/sisfokom.v14i4.2491.
- [10] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, "Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 11, no. 1, pp. 30–40, Mar. 2025, doi: 10.20473/jisebi.11.1.30-40.
- [11] B. Willie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint*

- Conference on Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
- [12] NxCode, “IndoBERT-Based Models for Indonesian NLP Tasks.”
- [13] L.-C. Cheng, Y.-L. Chen, and Y.-Y. Liao, “Aspect-based sentiment analysis with component focusing multi-head co-attention networks,” *Neurocomputing*, vol. 489, pp. 9–17, Jun. 2022, doi: 10.1016/j.neucom.2022.03.027.
- [14] B. Gledson, K. Rogage, A. Thompson, and H. Ponton, “Reporting on the Development of a Web-Based Prototype Dashboard for Construction Design Managers, Achieved through Design Science Research Methodology (DSRM),” *Buildings*, vol. 14, no. 2, p. 335, Jan. 2024, doi: 10.3390/buildings14020335.
- [15] L. Conrow, C. Fu, H. Huang, N. Andrienko, G. Andrienko, and R. Weibel, “A conceptual framework for developing dashboards for big mobility data,” *Cartogr. Geogr. Inf. Sci.*, vol. 50, no. 5, pp. 495–514, Sep. 2023, doi: 10.1080/15230406.2023.2190164.
- [16] M. C. Wuisang and G. Pangestu, “Aspect-Based Sentiment Analysis in Indonesian Hotel Reviews: Machine Learning versus Large Language Models,” in *2025 International Seminar on Application for Technology of Information and Communication (iSemantic)*, IEEE, Sep. 2025, pp. 444–449. doi: 10.1109/iSemantic67418.2025.11292470.
- [17] S. Studer *et al.*, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology,” *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, Apr. 2021, doi: 10.3390/make3020020.
- [18] A. Rahman and N. Hidayah, “Analisis Sentimen Berbasis Aspek pada Ulasan Pelanggan Menggunakan Pendekatan Deep Learning,” *Nusant. J. Sci. Technol.*, vol. 1, no. 2, pp. 45–52, 2024.
- [19] Bustamin, A. Prayogi, Siswanto, Rafrin, and Nurdin, “Text normalization for Indonesian slang words in sentiment analysis development,” *ICIC Express Lett. Part B Appl.*, vol. 16, no. 2, pp. 121–129, 2025.
- [20] Y. C. Hua, P. Denny, J. Wicker, and K. Taskova, “A systematic review of aspect-based sentiment analysis: domains, methods, and trends,” *Artif. Intell. Rev.*, vol. 57, no. 11, p. 296, Sep. 2024, doi: 10.1007/s10462-024-10906-z.
- [21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” Accessed: Sep. 01, 2026. [Online]. Available: <https://jmlr.org/papers/v15/srivastava14a.html>
- [22] Datasaur, “LLM Ranking and Benchmark for Sentiment Labeling Tasks.”
- [23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
- [24] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem. Medica*, pp. 276–282, 2012, doi: 10.11613/BM.2012.031.
- [25] Ilya Loshchilov and Frank Hutter, “Decoupled Weight Decay Regularization”.
- [26] N. Colmekcioglu, R. Marvi, P. Foroudi, and F. Okumus, “Generation, susceptibility, and response regarding negativity: An in-depth analysis on negative online reviews,” *J. Bus. Res.*, vol. 153, pp. 235–250, Dec. 2022, doi: 10.1016/j.jbusres.2022.08.033.
- [27] Miranda, “Analisis Sentimen Berbasis Aspek pada Ulasan Menggunakan Pendekatan Pembelajaran Mesin,” *Sist. J. Sist. Inf.*, vol. 11, no. 2, pp. 350–360, 2022.
- [28] A. Lindahl, “LLMs as annotators of argumentation,” in *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 242–252. doi: 10.18653/v1/2025.starsem-1.19.
- [29] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, Sep. 2023, doi: 10.1145/3560815.