

Comparative evaluation of deep learning models for aspect-level sentiment analysis using hybrid real and synthetic healthcare reviews

Eka Prima Ginting¹, Danny Mario Tua Panjaitan²

^{1,2}Universitas Sumatera Utara, Indonesia

Email: ekaprima@students.usu.ac.id; dannypanjaitan0@gmail.com

ABSTRAK

Analisis sentimen tingkat aspek memungkinkan identifikasi polaritas sentimen terhadap aspek tertentu sehingga memberikan informasi yang lebih terperinci dari ulasan layanan kesehatan. Penelitian ini bertujuan membandingkan performa LSTM, BiLSTM, CNN-BiLSTM, dan BERT dalam klasifikasi sentimen tingkat aspek. Dataset berupa korpus hibrida sebanyak 5.000 ulasan, terdiri atas 2.000 ulasan nyata dan 3.000 ulasan sintetis berbantuan AI, dengan lima aspek: kualitas pelayanan, tenaga medis, fasilitas, waktu tunggu, dan kebersihan. Dataset dibagi menjadi 80% data pelatihan dan 20% data pengujian. Performa model dievaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil menunjukkan BERT memperoleh performa tertinggi dengan *accuracy* 94,60%, *precision* 94,78%, *recall* 94,40%, dan *F1-score* 94,59%, diikuti CNN-BiLSTM, BiLSTM, dan LSTM. Aspek tenaga medis menghasilkan *accuracy* tertinggi sebesar 95,8%, sedangkan waktu tunggu terendah sebesar 92,7%. Hasil ini menunjukkan bahwa BERT memberikan performa terbaik di antara empat model yang dievaluasi pada kondisi eksperimen penelitian ini.

Kata Kunci: analisis sentimen tingkat aspek; BERT; *deep learning*; layanan kesehatan; data sintetis

ABSTRACT

Aspect-level sentiment analysis enables the identification of sentiment polarity toward specific aspects, providing more detailed information from healthcare reviews. This study aims to compare the performance of LSTM, BiLSTM, CNN-BiLSTM, and BERT for aspect-level sentiment classification. The dataset comprised a hybrid corpus of 5,000 reviews, including 2,000 real healthcare reviews and 3,000 AI-assisted synthetic reviews, covering five aspects: service quality, medical staff, facilities, waiting time, and cleanliness. The dataset was divided into 80% training data and 20% testing data. Model performance was evaluated using accuracy, precision, recall, and F1-score. The results showed that BERT achieved the highest performance, with 94.60% accuracy, 94.78% precision, 94.40% recall, and a 94.59% F1-score, followed by CNN-BiLSTM, BiLSTM, and LSTM. Medical staff achieved the highest aspect-level accuracy at 95.8%, whereas waiting time obtained the lowest at 92.7%. These findings indicate that BERT achieved the best classification performance among the four evaluated models under the experimental conditions of this study.

Keywords: aspect-level sentiment analysis; BERT; *deep learning*; healthcare reviews; synthetic data

Corresponding Author:

Eka Prima Ginting,
Universitas Sumatera Utara,
Jalan Dr. T. Mansur, Padang Bulan, Kecamatan Medan Baru, Kota Medan,
Provinsi Sumatera Utara, Indonesia
Email: ekaprima@students.usu.ac.id



1. PENDAHULUAN

Perkembangan platform komunikasi digital telah mengubah cara masyarakat menyampaikan pengalaman, opini, dan evaluasi terhadap berbagai layanan, termasuk layanan kesehatan. Ulasan yang disampaikan melalui platform digital dapat memuat informasi mengenai kualitas pelayanan, tenaga medis,

fasilitas, waktu tunggu, dan kebersihan. Informasi tersebut berpotensi menjadi sumber data tekstual yang bernilai untuk memahami persepsi pengguna terhadap pelayanan kesehatan. Namun, jumlah ulasan yang besar dan bentuknya yang tidak terstruktur menyebabkan analisis secara manual menjadi kurang efisien. Dalam konteks ini, *sentiment analysis* menjadi salah satu pendekatan dalam *Natural Language Processing* (NLP) yang dapat digunakan untuk mengidentifikasi dan menginterpretasikan opini yang terkandung dalam data tekstual (Cui et al., 2023; Hua et al., 2024).

Analisis sentimen secara umum mengklasifikasikan opini pada tingkat dokumen atau kalimat ke dalam polaritas tertentu, seperti positif, negatif, atau netral. Pendekatan tersebut dapat memberikan gambaran umum mengenai opini pengguna, tetapi belum tentu mampu menjelaskan sentimen terhadap atribut tertentu yang muncul dalam sebuah ulasan. Sebuah ulasan, misalnya, dapat memberikan penilaian positif terhadap tenaga medis tetapi pada saat yang sama mengandung penilaian negatif terhadap waktu tunggu. *Aspect-Based Sentiment Analysis* (ABSA) dikembangkan untuk memberikan analisis yang lebih terperinci dengan menghubungkan sentimen dengan aspek tertentu yang dibicarakan dalam teks. Dengan demikian, ABSA memungkinkan interpretasi opini yang lebih *fine-grained* dibandingkan analisis sentimen pada tingkat dokumen atau kalimat secara keseluruhan (Chen et al., 2024; Hua et al., 2024).

Perkembangan metode *deep learning* turut memperluas pendekatan yang dapat digunakan dalam analisis sentimen. Model *Long Short-Term Memory* (LSTM) memungkinkan pemodelan informasi sekuensial dan dependensi dalam teks, sedangkan *Bidirectional Long Short-Term Memory* (BiLSTM) memproses informasi dari dua arah untuk memperoleh representasi konteks yang lebih lengkap. Di sisi lain, *Convolutional Neural Network* (CNN) dapat digunakan untuk menangkap pola lokal dalam teks dan dapat dikombinasikan dengan BiLSTM untuk mengintegrasikan ekstraksi fitur lokal dengan pemodelan informasi sekuensial. Pendekatan hibrida CNN-BiLSTM telah digunakan dalam klasifikasi sentimen untuk memanfaatkan karakteristik kedua arsitektur tersebut (Ahmed & Wang, 2023; Elhassan et al., 2023).

Perkembangan berikutnya ditandai dengan penggunaan arsitektur berbasis *transformer*, seperti *Bidirectional Encoder Representations from Transformers* (BERT), yang membentuk representasi kontekstual berdasarkan hubungan antarkata dalam teks. Kemampuan tersebut menjadikan BERT dan pengembangannya banyak digunakan dalam penelitian analisis sentimen dan ABSA. Sejumlah penelitian juga telah mengembangkan pendekatan berbasis BERT maupun arsitektur hibrida untuk meningkatkan pemodelan konteks dan hubungan antara aspek dan sentimen (Alhadlaq & Altheneyan, 2024; Taye et al., 2025). Meskipun demikian, performa suatu arsitektur tetap dipengaruhi oleh karakteristik dataset, domain, konfigurasi eksperimen, serta kompleksitas ekspresi sentimen yang dianalisis. Oleh karena itu, hasil dari satu domain atau konfigurasi penelitian tidak secara langsung menunjukkan bahwa suatu arsitektur akan selalu lebih baik pada kondisi yang berbeda.

Selain pemilihan arsitektur, ketersediaan data merupakan persoalan penting dalam pengembangan model analisis sentimen. Dataset beranotasi dalam jumlah memadai dibutuhkan untuk mendukung proses pembelajaran model, sementara pengumpulan dan anotasi data tekstual dapat membutuhkan sumber daya yang cukup besar. Pendekatan *data augmentation* telah digunakan sebagai salah satu strategi untuk memperluas data tekstual dan meningkatkan variasi data yang tersedia untuk analisis sentimen (Gutiérrez Benítez et al., 2024). Dalam konteks penelitian ini, data sintetis berbasis AI digunakan sebagai pelengkap data ulasan kesehatan nyata, sehingga membentuk korpus hibrida yang terdiri atas data nyata dan sintetis. Penggunaan tersebut diposisikan sebagai strategi konstruksi dataset dan bukan sebagai bukti bahwa data sintetis secara independen meningkatkan performa atau kemampuan generalisasi model.

Penelitian sebelumnya menunjukkan perkembangan yang luas dalam analisis sentimen dengan menggunakan metode *machine learning*, *deep learning*, dan *transformer*. Braig et al. (2023), misalnya, mengevaluasi berbagai teknik *machine learning* untuk analisis sentimen pada data Twitter terkait COVID-19, sedangkan Ahmed dan Wang (2023) mengembangkan pendekatan *embedded-CNN* dengan BiLSTM untuk menganalisis sentimen pada ulasan produk. Pada pendekatan berbasis *transformer*, Taye et al. (2025) mengintegrasikan BERT dengan pendekatan berbasis ontologi untuk klasifikasi sentimen, sementara Alhadlaq dan Altheneyan (2024) mengembangkan arsitektur hibrida DistilRoBERTa dan *Graph Neural Network* untuk ABSA. Studi lainnya juga menunjukkan penerapan BERT dan model bahasa pada klasifikasi ulasan dan analisis sentimen di berbagai domain (Roumeliotis et al., 2024). Temuan-temuan tersebut menunjukkan bahwa pemodelan sentimen terus berkembang dari pendekatan sekuensial menuju representasi kontekstual yang semakin kompleks.

Meskipun demikian, masih terdapat ruang untuk mengevaluasi secara langsung arsitektur dengan karakteristik representasi yang berbeda dalam satu kondisi eksperimen yang sama, khususnya untuk klasifikasi sentimen pada aspek layanan kesehatan. Studi terdahulu menggunakan beragam dataset, domain, arsitektur, dan skenario evaluasi sehingga hasilnya tidak selalu dapat dibandingkan secara langsung. Selain itu, tantangan ABSA tidak hanya berkaitan dengan pemilihan model, tetapi juga dengan kompleksitas ekspresi sentimen pada aspek tertentu, termasuk ambiguitas, sentimen implisit, dan perbedaan tingkat kesulitan antar-aspek (Chifu &

Fournier, 2023; Hua et al., 2024). Oleh sebab itu, evaluasi beberapa arsitektur dalam dataset dan skema evaluasi yang sama dapat memberikan gambaran yang lebih terkontrol mengenai perbedaan performa model pada tugas klasifikasi sentimen tingkat aspek.

Berdasarkan permasalahan tersebut, penelitian ini melakukan evaluasi komparatif terhadap empat arsitektur *deep learning*, yaitu LSTM, BiLSTM, CNN-BiLSTM, dan BERT, untuk klasifikasi sentimen tingkat aspek pada ulasan layanan kesehatan. Dataset yang digunakan merupakan korpus hibrida berjumlah 5.000 ulasan, yang terdiri atas 2.000 ulasan kesehatan nyata dan 3.000 ulasan sintesis berbantuan AI. Analisis difokuskan pada lima aspek layanan kesehatan yang telah ditentukan, yaitu kualitas pelayanan (*service quality*), tenaga medis (*medical staff*), fasilitas (*facilities*), waktu tunggu (*waiting time*), dan kebersihan (*cleanliness*), dengan klasifikasi sentimen positif dan negatif. Performa keempat model dibandingkan menggunakan *accuracy*, *precision*, *recall*, dan *F1-score* sesuai dengan konfigurasi eksperimen yang digunakan dalam penelitian.

Kontribusi utama penelitian ini terletak pada evaluasi komparatif arsitektur dengan mekanisme representasi yang berbeda—mulai dari pemodelan sekuensial melalui LSTM, pemodelan sekuensial dua arah melalui BiLSTM, kombinasi ekstraksi fitur lokal dan sekuensial melalui CNN-BiLSTM, hingga representasi kontekstual berbasis *transformer* melalui BERT—dalam satu skenario klasifikasi sentimen tingkat aspek yang sama. Penggunaan kombinasi ulasan nyata dan sintesis memungkinkan pembentukan korpus yang lebih besar dan menyediakan variasi tekstual tambahan untuk eksperimen. Namun, karena penelitian ini tidak melakukan eksperimen pembandingan antara data nyata saja dan kombinasi data nyata-sintesis, kontribusi independen data sintesis terhadap peningkatan performa, *robustness*, atau generalisasi model tidak diklaim dalam penelitian ini.

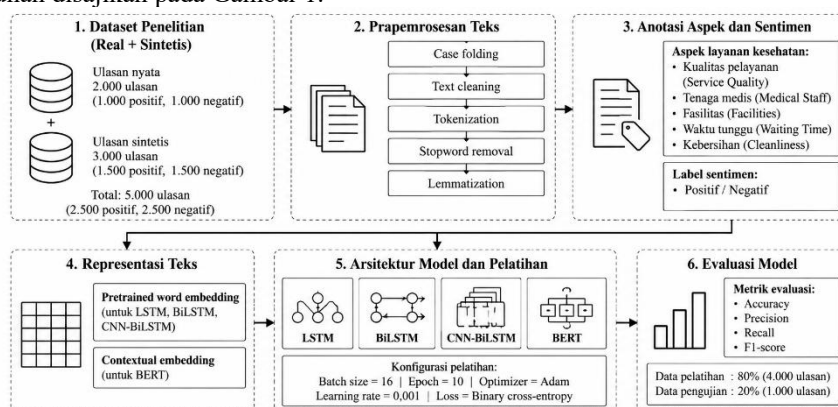
Dengan demikian, penelitian ini bertujuan untuk membandingkan performa LSTM, BiLSTM, CNN-BiLSTM, dan BERT dalam mengklasifikasikan sentimen positif dan negatif pada lima aspek layanan kesehatan menggunakan korpus hibrida ulasan nyata dan sintesis. Evaluasi tersebut digunakan untuk mengidentifikasi perbedaan performa keempat arsitektur pada kondisi eksperimen yang dilaporkan serta memberikan gambaran mengenai karakteristik model yang lebih sesuai untuk klasifikasi sentimen tingkat aspek pada data ulasan layanan kesehatan.

2. METODE PENELITIAN

A. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental untuk membandingkan performa empat arsitektur *deep learning* dalam klasifikasi sentimen tingkat aspek pada ulasan layanan kesehatan. Arsitektur yang dievaluasi meliputi *Long Short-Term Memory* (LSTM), *Bidirectional Long Short-Term Memory* (BiLSTM), *Convolutional Neural Network–Bidirectional Long Short-Term Memory* (CNN-BiLSTM), dan *Bidirectional Encoder Representations from Transformers* (BERT).

Penelitian dilaksanakan melalui beberapa tahapan yang meliputi penyusunan korpus ulasan kesehatan nyata dan sintesis, prapemrosesan teks, anotasi aspek dan sentimen, pembentukan representasi teks, pelatihan empat arsitektur, dan evaluasi performa klasifikasi. Seluruh model dievaluasi pada skema pembagian data yang sama untuk memperoleh perbandingan performa dalam kondisi eksperimen yang dilaporkan. Alur penelitian secara keseluruhan disajikan pada Gambar 1.



Gambar 1. Alur penelitian klasifikasi sentimen tingkat aspek

B. Dataset Penelitian

Dataset penelitian terdiri atas 5.000 ulasan terkait layanan kesehatan, yang merupakan kombinasi 2.000 ulasan nyata dan 3.000 ulasan sintesis berbantuan AI. Dataset memuat dua kelas sentimen, yaitu positif dan negatif, serta mencakup lima aspek layanan kesehatan yang telah ditentukan, yaitu kualitas pelayanan

(*service quality*), tenaga medis (*medical staff*), fasilitas (*facilities*), waktu tunggu (*waiting time*), dan kebersihan (*cleanliness*).

Komposisi keseluruhan dataset dirangkum pada Tabel 1.

Tabel 1. Komposisi dataset penelitian

Sumber data	Positif	Negatif	Total
Ulasan nyata	1.000	1.000	2.000
Ulasan sintetis	1.500	1.500	3.000
Total	2.500	2.500	5.000

Komposisi tersebut menunjukkan bahwa dataset memiliki distribusi kelas sentimen yang seimbang, dengan masing-masing 2.500 ulasan positif dan 2.500 ulasan negatif. Dataset kemudian dibagi dengan rasio 80:20, yaitu 4.000 ulasan sebagai data pelatihan dan 1.000 ulasan sebagai data pengujian. Pembagian tersebut mengikuti konfigurasi eksperimen pada penelitian ini.

1) Ulasan Kesehatan Nyata

Subset data nyata terdiri atas 2.000 ulasan, masing-masing 1.000 ulasan positif dan 1.000 ulasan negatif. Data berasal dari ulasan layanan kesehatan yang tersedia secara publik dan memuat opini mengenai pelayanan rumah sakit, tenaga medis, waktu tunggu, fasilitas, dan kebersihan.

Pada tahap penyaringan, entri duplikat, kalimat yang tidak lengkap, dan teks yang tidak relevan dikeluarkan dari dataset. Proses tersebut dilakukan untuk mempertahankan ulasan yang memiliki konteks yang sesuai dengan aspek layanan kesehatan yang dianalisis.

Contoh data nyata ditunjukkan pada Tabel 2.

Tabel 2. Contoh ulasan kesehatan nyata

Ulasan	Aspek	Sentimen
"The doctor explained the treatment clearly and professionally."	Medical Staff	Positive
"The waiting time for consultation was extremely long."	Waiting Time	Negative
"The hospital facilities were modern and well maintained."	Facilities	Positive
"The administrative service was very slow and inefficient."	Service Quality	Negative
"The nurses were friendly and responsive during treatment."	Medical Staff	Positive

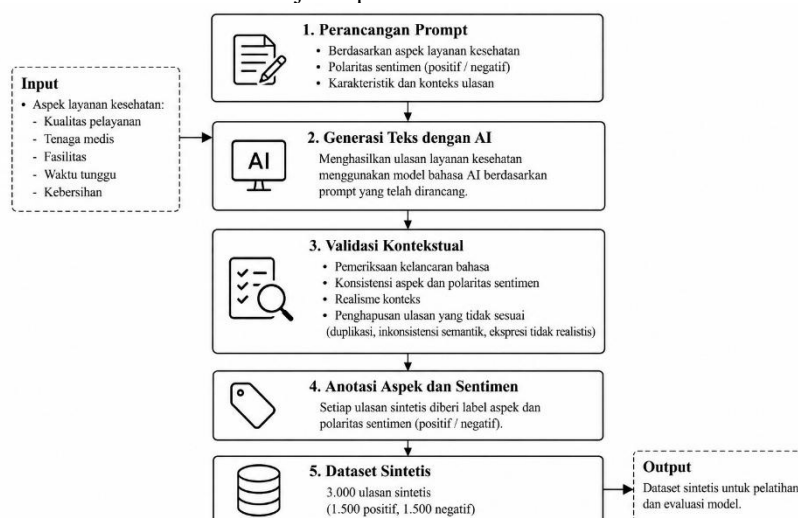
C. Pembentukan Data Sintetis

Subset sintetis terdiri atas 3.000 ulasan, dengan distribusi seimbang berupa 1.500 ulasan positif dan 1.500 ulasan negatif. Data sintetis dibentuk melalui pendekatan pembuatan teks berbantuan AI dengan *prompt* yang diarahkan berdasarkan aspek layanan kesehatan, polaritas sentimen, dan konteks ulasan.

Proses pembentukan data menerapkan beberapa batasan untuk menjaga kesesuaian data yang dihasilkan, yaitu: (1) konsistensi aspek, (2) konsistensi polaritas sentimen, (3) variasi linguistik, (4) realisme konteks, dan (5) variasi struktur kalimat. Ulasan sintetis yang dihasilkan kemudian diperiksa secara manual untuk mengeluarkan struktur yang berulang, opini yang tidak konsisten secara semantik, serta ekspresi yang dinilai tidak sesuai dengan konteks layanan kesehatan.

Dengan demikian, data sintetis dalam penelitian ini digunakan sebagai pelengkap data nyata dalam pembentukan korpus hibrida. Penelitian ini tidak melakukan eksperimen *ablation* yang membandingkan pelatihan menggunakan data nyata saja dengan pelatihan menggunakan kombinasi data nyata dan sintetis. Oleh karena itu, prosedur ini tidak digunakan sebagai dasar untuk menyimpulkan bahwa penambahan data sintetis secara independen meningkatkan performa atau generalisasi model.

Alur pembentukan data sintetis disajikan pada Gambar 2.



Gambar 2. Alur pembentukan data ulasan sintetis

(Eka Prima Ginting)

Contoh data sintetis ditunjukkan pada Tabel 3.

Tabel 3. Contoh ulasan kesehatan sintetis

Ulasan sintetis	Aspek	Sentimen
"The medical staff were highly professional, although the waiting process was frustratingly slow."	Waiting Time	Negative
"The hospital environment felt comfortable and hygienic during my visit."	Cleanliness	Positive
"The nurses provided excellent assistance throughout the treatment."	Medical Staff	Positive
"The consultation scheduling system was extremely inefficient."	Service Quality	Negative
"The healthcare facilities were advanced and technologically modern."	Facilities	Positive

D. Prapemrosesan Teks

Prapemrosesan dilakukan sebelum data digunakan dalam proses pelatihan model. Berdasarkan prosedur penelitian, tahap prapemrosesan meliputi case folding, text cleaning, tokenization, stopword removal, dan lemmatization.

Pada tahap *case folding*, seluruh karakter dalam teks diubah menjadi huruf kecil untuk menyeragamkan bentuk kata. *Text cleaning* dilakukan dengan menghilangkan karakter khusus, tanda baca, *hyperlink*, angka, dan spasi berlebih. Selanjutnya, *tokenization* digunakan untuk memisahkan teks menjadi unit token. *Stopword removal* dilakukan dengan menghapus kata-kata yang memiliki kontribusi semantik terbatas, sedangkan *lemmatization* digunakan untuk mengubah kata ke bentuk dasarnya.

Secara umum, transformasi prapemrosesan direpresentasikan sebagai:

$$d' = P(d) \quad (1)$$

dengan d merupakan dokumen teks asli, $P(\cdot)$ merupakan fungsi prapemrosesan, dan d' merupakan teks hasil normalisasi.

E. Anotasi Aspek dan Sentimen

Penelitian menggunakan lima aspek layanan kesehatan yang telah ditentukan sebelumnya, yaitu kualitas pelayanan, tenaga medis, fasilitas, waktu tunggu, dan kebersihan. Setiap ulasan dikaitkan dengan aspek yang relevan dan diberikan label sentimen positif atau negatif.

Anotasi dilakukan secara manual dan semi-otomatis untuk menjaga konsistensi aspek dan polaritas sentimen. Namun, informasi mengenai jumlah anotator, mekanisme penyelesaian perbedaan anotasi, dan ukuran *inter-annotator agreement* tidak terdokumentasi sehingga tingkat kesepakatan antar-anotator tidak dapat dilaporkan.

F. Representasi Teks dan Arsitektur Model

Representasi masukan disesuaikan dengan arsitektur yang dievaluasi. LSTM, BiLSTM, dan CNN-BiLSTM menggunakan representasi vektor kata berbasis *pretrained word embedding*, sedangkan BERT menggunakan representasi kontekstual melalui mekanisme *bidirectional encoding*.

Untuk BERT, representasi kontekstual secara umum dapat dinyatakan sebagai:

$$h_i = \text{BERT}(w_1, w_2, \dots, w_n) \quad (2)$$

dengan h_i merupakan representasi kontekstual dan $w_1, w_2, \dots, w_n$ merupakan token masukan.

Empat arsitektur yang dibandingkan adalah sebagai berikut. LSTM digunakan untuk memodelkan dependensi sekuensial dalam teks. BiLSTM memperluas pemrosesan tersebut dengan membaca urutan dalam arah maju dan mundur. CNN-BiLSTM mengombinasikan ekstraksi pola lokal oleh CNN dengan pemodelan sekuensial dua arah melalui BiLSTM. Sementara itu, BERT menggunakan arsitektur *transformer* dan mekanisme *self-attention* untuk membentuk representasi kontekstual.

Mekanisme *scaled dot-product attention* secara umum dinyatakan sebagai:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Dengan Q , K , dan V masing-masing merupakan matriks *query*, *key*, dan *value*, sedangkan d_k merupakan dimensi *key*.

G. Pelatihan dan Konfigurasi Eksperimen

Dataset dibagi menjadi data pelatihan dan data pengujian dengan rasio 80:20. Sebanyak 4.000 ulasan digunakan dalam proses pelatihan, sedangkan 1.000 ulasan digunakan sebagai data pengujian.

Konfigurasi eksperimen yang digunakan disajikan pada Tabel 4.

Tabel 4. Konfigurasi eksperimen

Parameter	Nilai
Ukuran dataset	5.000 ulasan
Data pelatihan	80% (4.000 ulasan)
Data pengujian	20% (1.000 ulasan)
Batch size	16
Epoch	10

Parameter	Nilai
Optimizer	Adam
Learning rate	0,001
Loss	Binary cross-entropy
Metrik evaluasi	Accuracy, Precision, Recall, F1-score

Fungsi *binary cross-entropy* dinyatakan sebagai:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (4)$$

dengan y_i merupakan label aktual, $\hat{y}_i$ merupakan probabilitas prediksi, dan N merupakan jumlah sampel.

H. Evaluasi Model

Performa model dievaluasi menggunakan accuracy, precision, recall, dan F1-score berdasarkan nilai *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN).

Accuracy dihitung menggunakan:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision dihitung menggunakan:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall dihitung menggunakan:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

dan *F1-score* dihitung menggunakan:

$$F1 = 2 \left(\frac{Precision \times Recall}{Precision + Recall} \right) \quad (8)$$

Keempat metrik tersebut digunakan untuk membandingkan performa LSTM, BiLSTM, CNN-BiLSTM, dan BERT pada data pengujian yang sama. *Confusion matrix* selanjutnya digunakan untuk memberikan analisis tambahan terhadap pola klasifikasi model dengan performa terbaik.

3. HASIL DAN PEMBAHASAN

A. Gambaran Eksperimen

Eksperimen dilakukan menggunakan korpus hibrida yang terdiri atas 5.000 ulasan layanan kesehatan, yaitu 2.000 ulasan nyata dan 3.000 ulasan sintetis berbantuan AI. Dataset memiliki distribusi kelas yang seimbang, yaitu 2.500 ulasan positif dan 2.500 ulasan negatif. Analisis difokuskan pada lima aspek layanan kesehatan yang telah ditentukan, yaitu kualitas pelayanan (*service quality*), tenaga medis (*medical staff*), fasilitas (*facilities*), waktu tunggu (*waiting time*), dan kebersihan (*cleanliness*).

Dataset dibagi dengan rasio 80:20 sehingga 4.000 ulasan digunakan sebagai data pelatihan dan 1.000 ulasan sebagai data pengujian. Empat arsitektur yang dibandingkan adalah LSTM, BiLSTM, CNN-BiLSTM, dan BERT. Performa klasifikasi dievaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*.

B. Perbandingan Performa Model

Hasil evaluasi menunjukkan adanya peningkatan performa dari LSTM, BiLSTM, dan CNN-BiLSTM menuju BERT. Hasil lengkap berdasarkan nilai yang dilaporkan dalam eksperimen disajikan pada Tabel 5.

Tabel 5. Perbandingan performa model

Model	Accuracy	Precision	Recall	F1-score
LSTM	87,4%	86,9%	86,1%	86,5%
BiLSTM	89,2%	88,6%	88,1%	88,3%
CNN-BiLSTM	91,1%	90,5%	90,2%	90,3%
BERT	94,60%	94,78%	94,40%	94,59%

LSTM memperoleh *accuracy* sebesar 87,4% dan *F1-score* sebesar 86,5%, yang merupakan performa terendah di antara empat model. Penggunaan pemrosesan dua arah pada BiLSTM diikuti peningkatan *accuracy* menjadi 89,2% dan *F1-score* menjadi 88,3%. CNN-BiLSTM menghasilkan performa lebih tinggi dengan *accuracy* 91,1% dan *F1-score* 90,3%. Sementara itu, BERT memperoleh performa tertinggi dengan *accuracy* sebesar 94,60%, *precision* 94,78%, *recall* 94,40%, dan *F1-score* 94,59%. Nilai tersebut dihitung berdasarkan confusion matrix hasil klasifikasi terhadap 1.000 data pengujian.

Jika dibandingkan berdasarkan *accuracy*, BiLSTM berada 1,8 poin persentase di atas LSTM. CNN-BiLSTM selanjutnya meningkat 1,9 poin persentase dibandingkan BiLSTM, sedangkan BERT berada 3,5 poin persentase di atas CNN-BiLSTM dan 7,2 poin persentase di atas LSTM. Pola yang serupa juga terlihat pada *precision*, *recall*, dan *F1-score* yang dilaporkan.

Hasil tersebut menunjukkan bahwa dalam kondisi eksperimen penelitian ini, BERT memberikan performa klasifikasi tertinggi di antara empat arsitektur yang dievaluasi. Temuan tersebut dapat dikaitkan

(Eka Prima Ginting)

dengan perbedaan mekanisme representasi teks pada masing-masing model. LSTM memproses informasi secara sekuensial, sedangkan BiLSTM menggunakan informasi dari arah maju dan mundur. CNN-BiLSTM mengombinasikan ekstraksi pola lokal oleh CNN dengan pemodelan sekuensial dua arah. BERT berbeda karena menggunakan representasi kontekstual berbasis *transformer* dan mekanisme *self-attention*.

Kecenderungan tersebut sejalan dengan perkembangan penelitian ABSA yang semakin banyak menggunakan representasi kontekstual dan arsitektur berbasis *transformer* untuk menangkap hubungan antara konteks, aspek, dan polaritas sentimen (Alhadlaq & Altheneyan, 2024; Chen et al., 2024; Taye et al., 2025). Meskipun demikian, hasil penelitian ini tidak dimaksudkan untuk menyimpulkan bahwa BERT secara universal lebih unggul daripada seluruh arsitektur recurrent atau hybrid. Kesimpulan dibatasi pada empat model, dataset, dan konfigurasi eksperimen yang digunakan dalam penelitian ini.

C. Analisis Confusion Matrix BERT

Karena BERT memperoleh *accuracy* tertinggi, analisis lebih lanjut dilakukan menggunakan *confusion matrix*. Hasil klasifikasi terhadap 1.000 data pengujian disajikan pada Tabel 6.

Tabel 6. Confusion matrix model BERT

Aktual / Prediksi	Positif	Negatif
Positif	472	28
Negatif	26	474

BERT mengklasifikasikan dengan benar 472 dari 500 ulasan positif dan 474 dari 500 ulasan negatif. Sebanyak 28 ulasan positif diklasifikasikan sebagai negatif dan 26 ulasan negatif diklasifikasikan sebagai positif. Berdasarkan confusion matrix tersebut, diperoleh *accuracy* sebesar 94,60%, *precision* 94,78%, *recall* 94,40%, dan *F1-score* 94,59%.

D. Performa Berdasarkan Aspek Layanan Kesehatan

Evaluasi berikutnya membandingkan performa klasifikasi pada lima aspek layanan kesehatan. Hasil yang dilaporkan disajikan pada Tabel 7.

Tabel 7. Performa klasifikasi berdasarkan aspek

Aspek	Accuracy	Precision	Recall	F1-score
Service Quality	95,2%	94,8%	94,5%	94,7%
Medical Staff	95,8%	95,3%	95,0%	95,1%
Facilities	94,3%	93,9%	93,6%	93,8%
Waiting Time	92,7%	92,3%	91,9%	92,1%
Cleanliness	94,6%	94,1%	93,8%	94,0%

Aspek Medical Staff memperoleh performa tertinggi dengan *accuracy* 95,8% dan *F1-score* 95,1%. Sebaliknya, Waiting Time memperoleh nilai terendah dengan *accuracy* 92,7% dan *F1-score* 92,1%. Selisih *accuracy* antara kedua aspek tersebut sebesar 3,1 poin persentase, menunjukkan bahwa tingkat keberhasilan klasifikasi tidak identik pada seluruh aspek layanan.

Salah satu interpretasi yang mungkin adalah perbedaan karakteristik linguistik pada setiap aspek. Ulasan mengenai tenaga medis dapat mengandung ekspresi evaluatif yang relatif eksplisit, seperti *professional*, *friendly*, *helpful*, *rude*, atau *unresponsive*. Sebaliknya, ketidakpuasan terhadap waktu tunggu dapat diekspresikan secara lebih implisit. Sebagai contoh, kalimat “*I had to wait longer than expected before the consultation*” tidak mengandung kata negatif yang sangat eksplisit, tetapi tetap dapat mengindikasikan ketidakpuasan terhadap waktu tunggu.

Interpretasi tersebut konsisten dengan gagasan bahwa tingkat kesulitan dalam ABSA dapat dipengaruhi oleh cara sentimen diekspresikan terhadap suatu aspek (Chifu & Fournier, 2023). Namun, karena penelitian ini tidak melakukan analisis kesalahan linguistik secara terpisah dan sistematis untuk setiap aspek, karakteristik tersebut harus dipahami sebagai interpretasi yang mungkin menjelaskan pola hasil, bukan sebagai penyebab yang telah dibuktikan secara empiris.

E. Kompleksitas Sentimen pada Ulasan Multi-Aspek

Salah satu tantangan dalam klasifikasi sentimen tingkat aspek adalah keberadaan beberapa aspek dan polaritas sentimen dalam satu ulasan. Sebagai ilustrasi, perhatikan ulasan berikut:

“*The nurses were very professional, but the administrative process was frustrating.*”

Kalimat tersebut memuat sentimen positif terhadap Medical Staff melalui ekspresi “*very professional*”, tetapi pada saat yang sama memuat sentimen negatif terhadap Service Quality melalui ekspresi “*administrative process was frustrating*”.

Contoh tersebut menunjukkan mengapa klasifikasi tingkat aspek berbeda dari klasifikasi sentimen umum. Model tidak hanya perlu mengenali polaritas kalimat secara keseluruhan, tetapi juga menghubungkan ekspresi sentimen dengan aspek yang sesuai. Ambiguitas, sentimen implisit, dan ekspresi multi-aspek merupakan bagian dari kesulitan yang telah dibahas dalam penelitian ABSA (Chifu & Fournier, 2023; Hua et al., 2024).

Meskipun demikian, penelitian ini tidak melakukan kategorisasi kuantitatif terhadap 54 kesalahan BERT berdasarkan jenis kesalahan linguistik. Oleh karena itu, tidak dapat disimpulkan bahwa ulasan multi-aspek merupakan penyebab utama kesalahan BERT. Contoh tersebut digunakan untuk menjelaskan salah satu bentuk kompleksitas linguistik yang relevan terhadap tugas yang dievaluasi.

F. Waktu Pelatihan dan Trade-off Performa-Komputasi

Selain performa klasifikasi, penelitian juga mencatat waktu pelatihan masing-masing model selama 10 *epoch*. Hasilnya ditunjukkan pada Tabel 8.

Tabel 8. Waktu pelatihan model

Model	Epoch	Waktu pelatihan
LSTM	10	14 menit
BiLSTM	10	19 menit
CNN-BiLSTM	10	24 menit
BERT	10	41 menit

Data eksperimen menunjukkan bahwa LSTM membutuhkan waktu pelatihan terpendek, yaitu 14 menit, sedangkan BERT membutuhkan waktu terpanjang, yaitu 41 menit. BERT dengan demikian membutuhkan sekitar 2,93 kali waktu pelatihan LSTM.

Jika waktu pelatihan dibandingkan dengan *accuracy*, terlihat adanya *trade-off*. LSTM memperoleh *accuracy* 87,4% dengan waktu pelatihan 14 menit, sedangkan CNN-BiLSTM mencapai 91,1% dalam 24 menit. BERT memberikan *accuracy* tertinggi, yaitu 94,6%, tetapi membutuhkan waktu pelatihan paling lama, yaitu 41 menit.

Temuan tersebut menunjukkan bahwa pemilihan model tidak hanya dapat mempertimbangkan performa klasifikasi. Dalam kondisi komputasi yang serupa dengan eksperimen ini, BERT menawarkan *accuracy* tertinggi tetapi dengan kebutuhan waktu pelatihan yang lebih besar. CNN-BiLSTM berada di antara BERT dan model recurrent lainnya dari sisi performa dan waktu pelatihan.

G. Posisi Data Sintetis dalam Hasil Penelitian

Korpus penelitian menggabungkan 2.000 ulasan nyata dan 3.000 ulasan sintetis, sehingga penggunaan data sintetis memungkinkan dataset eksperimen mencapai 5.000 ulasan dengan distribusi sentimen yang seimbang. Data sintetis juga dirancang untuk menyediakan variasi struktur kalimat, aspek, polaritas, dan konteks sesuai dengan prosedur pembentukan data yang telah dijelaskan pada bagian Metode.

Namun, hasil tersebut tidak dapat digunakan untuk menyimpulkan bahwa data sintetis meningkatkan *accuracy*, F1-score, robustness, atau generalisasi model. Penelitian tidak membandingkan skenario:

Real Only dengan Real + Synthetic

Seluruh model dilatih dan dievaluasi menggunakan skema korpus hibrida yang sama. Dengan demikian, kontribusi yang dapat dinyatakan berdasarkan desain penelitian adalah bahwa data sintetis menambah ukuran dan variasi yang dirancang dalam korpus, bukan bahwa data sintetis secara kausal meningkatkan performa model.

H. Implikasi Temuan

Secara keseluruhan, hasil penelitian menunjukkan tiga temuan utama. Pertama, BERT memperoleh performa tertinggi di antara empat model pada konfigurasi eksperimen yang digunakan, dengan *accuracy* 94,60%. Kedua, performa berbeda antar-aspek; Medical Staff memberikan *accuracy* tertinggi sebesar 95,8%, sedangkan Waiting Time menghasilkan nilai terendah sebesar 92,7%. Ketiga, peningkatan performa BERT disertai waktu pelatihan yang lebih panjang dibandingkan tiga arsitektur lainnya.

Dari sisi penerapan, klasifikasi sentimen tingkat aspek berpotensi memberikan informasi yang lebih terperinci dibandingkan penilaian sentimen keseluruhan karena opini dapat dikaitkan dengan dimensi pelayanan tertentu. Dalam konteks layanan kesehatan, informasi tersebut dapat digunakan untuk mengidentifikasi apakah opini positif atau negatif berkaitan dengan kualitas pelayanan, tenaga medis, fasilitas, waktu tunggu, atau kebersihan. Potensi tersebut mendukung penggunaan analisis sentimen sebagai alat bantu analitik terhadap umpan balik pengguna, tetapi implementasinya pada lingkungan nyata tetap memerlukan validasi lebih lanjut menggunakan data yang independen dan representatif.

I. Keterbatasan Penelitian

Interpretasi hasil penelitian perlu mempertimbangkan beberapa keterbatasan. Pertama, klasifikasi hanya menggunakan dua kelas sentimen, yaitu positif dan negatif, sehingga opini netral atau ekspresi dengan polaritas yang tidak jelas belum direpresentasikan. Kedua, sebanyak 3.000 dari 5.000 ulasan merupakan data sintetis; data tersebut mungkin belum sepenuhnya merepresentasikan variasi linguistik dan pengalaman pasien pada kondisi nyata.

Ketiga, penelitian tidak menggunakan eksperimen real-only versus real+synthetic, sehingga pengaruh independen data sintetis terhadap performa dan generalisasi tidak dapat ditentukan. Keempat, evaluasi didasarkan pada satu pembagian train-test 80:20 dan penelitian tidak melaporkan *repeated runs*, variasi *random seed*, maupun *cross-validation*. Dengan demikian, variasi performa antar-pelatihan belum dapat dievaluasi.

Kelima, tidak terdapat external validation dataset, sehingga kemampuan generalisasi hasil terhadap sumber ulasan kesehatan yang berbeda belum diuji. Keenam, informasi rinci mengenai sumber 2.000 ulasan nyata dan konfigurasi model AI yang menghasilkan data sintesis tidak terdokumentasi secara lengkap. Ketujuh, beberapa konfigurasi arsitektur—seperti jumlah unit recurrent, ukuran *filter* CNN, *dropout rate*, panjang sekuens, dimensi *embedding*, dan versi/checkpoint BERT—juga tidak tersedia secara lengkap dalam catatan eksperimen.

Kedelapan, proses anotasi disebut dilakukan secara manual dan semi-otomatis, tetapi jumlah anotator dan ukuran *inter-annotator agreement* tidak terdokumentasi. Kesembilan, penelitian hanya menggunakan informasi tekstual dan belum memasukkan sumber informasi multimodal seperti rating, gambar, atau metadata. Pengembangan ke arah multimodal merupakan salah satu arah yang berkembang dalam penelitian analisis sentimen (Singh et al., 2024; Weng et al., 2024).

Keterbatasan tersebut membatasi cakupan generalisasi hasil, tetapi tidak mengubah hasil komparatif yang dilaporkan untuk kondisi eksperimen penelitian ini. Penelitian selanjutnya dapat menguji kelas sentimen netral, dataset multilingual, validasi eksternal, *explainable AI*, arsitektur *lightweight transformer*, serta desain *ablation* yang secara khusus membandingkan penggunaan data nyata dengan kombinasi data nyata dan sintesis.

4. KESIMPULAN

Penelitian ini membandingkan performa empat arsitektur *deep learning*, yaitu LSTM, BiLSTM, CNN-BiLSTM, dan BERT, untuk klasifikasi sentimen tingkat aspek pada ulasan layanan kesehatan. Eksperimen menggunakan korpus hibrida sebanyak 5.000 ulasan yang terdiri atas 2.000 ulasan nyata dan 3.000 ulasan sintesis berbantuan AI, dengan dua kelas sentimen, yaitu positif dan negatif. Analisis dilakukan terhadap lima aspek layanan kesehatan yang telah ditentukan, meliputi kualitas pelayanan (*service quality*), tenaga medis (*medical staff*), fasilitas (*facilities*), waktu tunggu (*waiting time*), dan kebersihan (*cleanliness*).

Berdasarkan hasil eksperimen yang dilaporkan, BERT memperoleh performa tertinggi di antara empat model yang dibandingkan, dengan *accuracy* sebesar 94,60% dan *F1-score* sebesar 94,59%, diikuti CNN-BiLSTM, BiLSTM, dan LSTM. Hasil tersebut menunjukkan bahwa, dalam kondisi eksperimen penelitian ini, representasi kontekstual berbasis *transformer* pada BERT memberikan performa klasifikasi yang lebih tinggi dibandingkan arsitektur recurrent dan hybrid yang dievaluasi. CNN-BiLSTM menempati posisi kedua dengan *accuracy* 91,1% dan *F1-score* 90,3%, yang menunjukkan bahwa kombinasi ekstraksi fitur lokal dan pemodelan sekuensial juga memberikan hasil yang kompetitif.

Evaluasi berdasarkan aspek menunjukkan adanya variasi performa klasifikasi. Aspek Medical Staff menghasilkan performa tertinggi dengan *accuracy* 95,8% dan *F1-score* 95,1%, sedangkan Waiting Time memperoleh performa terendah dengan *accuracy* 92,7% dan *F1-score* 92,1%. Temuan ini menunjukkan bahwa performa klasifikasi tidak seragam pada seluruh aspek layanan kesehatan dan dapat berkaitan dengan karakteristik ekspresi sentimen yang terdapat pada masing-masing aspek.

Penggunaan 3.000 ulasan sintesis memungkinkan pembentukan korpus hibrida yang lebih besar dan menyediakan variasi tekstual tambahan sebagai pelengkap 2.000 ulasan nyata. Namun, penelitian ini tidak melakukan perbandingan antara skenario data nyata saja dan kombinasi data nyata-sintesis. Oleh karena itu, hasil penelitian tidak dapat digunakan untuk menyimpulkan bahwa penggunaan data sintesis secara independen meningkatkan performa, *robustness*, atau kemampuan generalisasi model. Data sintesis dalam penelitian ini lebih tepat diposisikan sebagai strategi pelengkap dalam konstruksi dataset.

Penelitian ini masih memiliki keterbatasan, terutama penggunaan dua kelas sentimen, satu pembagian data pelatihan-pengujian 80:20, belum adanya validasi eksternal dan eksperimen berulang, serta keterbatasan dokumentasi mengenai sumber data nyata, konfigurasi pembentukan data sintesis, proses anotasi, dan beberapa konfigurasi spesifik arsitektur. Penelitian selanjutnya dapat mengembangkan evaluasi dengan memasukkan kelas netral, data multilingual, validasi pada dataset eksternal, eksperimen *ablation* untuk mengukur kontribusi data sintesis secara khusus, metode *explainable AI*, serta arsitektur *lightweight transformer*. Dengan mempertimbangkan batasan tersebut, hasil penelitian ini memberikan bukti komparatif dalam ruang lingkup eksperimen yang dilakukan bahwa BERT merupakan model dengan performa tertinggi di antara empat arsitektur yang dievaluasi untuk klasifikasi sentimen tingkat aspek pada korpus hibrida ulasan layanan kesehatan.

REFERENSI

- Ahamad, R., & Mishra, K. N. (2025). Exploring sentiment analysis in handwritten and E-text documents using advanced machine learning techniques: A novel approach. *Journal of Big Data*, 12(1), 11. <https://doi.org/10.1186/s40537-025-01064-2>
- Ahmed, Z., & Wang, J. (2023). A fine-grained deep learning model using embedded-CNN with BiLSTM for exploiting product sentiments. *Alexandria Engineering Journal*, 65, 731–747. <https://doi.org/10.1016/j.aej.2022.10.037>

- Alhadlaq, A., & Altheneyan, A. (2024). Distilroberta2gnn: A new hybrid deep learning approach for aspect-based sentiment analysis. *PeerJ Computer Science*, 10, e2267. <https://doi.org/10.7717/peerj-cs.2267>
- Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, 56(11), 12505–12560. <https://doi.org/10.1007/s10462-023-10442-2>
- Braig, N., Benz, A., Voth, S., Breitenbach, J., & Buettner, R. (2023). Machine learning techniques for sentiment analysis of COVID-19-related Twitter data. *IEEE Access*, 11, 14778–14803. <https://doi.org/10.1109/ACCESS.2023.3242234>
- Chen, X., Xie, H., Qin, S. J., Chai, Y., Tao, X., & Wang, F. L. (2024). Cognitive-inspired deep learning models for aspect-based sentiment analysis: A retrospective overview and bibliometric analysis. *Cognitive Computation*, 16(6), 3518–3556. <https://doi.org/10.1007/s12559-024-10331-y>
- Chifu, A.-G., & Fournier, S. (2023). Sentiment difficulty in aspect-based sentiment analysis. *Mathematics*, 11(22), 4647. <https://doi.org/10.3390/math11224647>
- Cui, J., Wang, Z., Ho, S.-B., & Cambria, E. (2023). Survey on sentiment analysis: Evolution of research methods and topics. *Artificial Intelligence Review*, 56(8), 8469–8510. <https://doi.org/10.1007/s10462-022-10386-z>
- Elhassan, N., Varone, G., Ahmed, R., Gogate, M., Dashtipour, K., Almoamari, H., El-Affendi, M. A., Al-Tamimi, B. N., Albalwy, F., & Hussain, A. (2023). Arabic sentiment analysis based on word embeddings and deep learning. *Computers*, 12(6), 126. <https://doi.org/10.3390/computers12060126>
- Gutiérrez Benítez, R., Segura Navarrete, A., Vidal-Castro, C., & Martínez-Araneda, C. (2024). Guide for the application of the data augmentation approach on sets of texts in Spanish for sentiment and emotion analysis. *PLOS ONE*, 19(9), e0310707. <https://doi.org/10.1371/journal.pone.0310707>
- Hua, Y. C., Denny, P., Wicker, J., & Taskova, K. (2024). A systematic review of aspect-based sentiment analysis: Domains, methods, and trends. *Artificial Intelligence Review*, 57(11), 296. <https://doi.org/10.1007/s10462-024-10906-z>
- Rehman, S. U., Blessing, O. O., & Ali, A. (2025). A comparative study of X data about the NHS using sentiment analysis. *Big Data and Cognitive Computing*, 9(10), 244. <https://doi.org/10.3390/bdcc9100244>
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Leveraging large language models in tourism: A comparative study of the latest GPT Omni models and BERT NLP for customer review classification and sentiment analysis. *Information*, 15(12), 792. <https://doi.org/10.3390/info15120792>
- Singh, U., Abhishek, K., & Azad, H. K. (2024). A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56(9), 1–38. <https://doi.org/10.1145/3652149>
- Taye, M. M., Abulail, R., Al-Ifan, B., & Alsuhimat, F. (2025). Enhanced sentiment classification through ontology-based sentiment analysis with BERT. *Journal of Internet Services and Information Security*, 15(1), 236–256. <https://doi.org/10.58346/JISIS.2025.11.015>
- Vamsidhar, D., Desai, P., Shahade, A. K., Patil, S., & Deshmukh, P. V. (2025). Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. *Scientific Reports*, 15(1), 25440. <https://doi.org/10.1038/s41598-025-09000-3>
- Weng, Y., Chen, L., Wang, S., Ye, X., Liu, X., & Liu, Z. (2024). MIECF: Multi-faceted information extraction and cross-mixture fusion for multimodal aspect-based sentiment analysis. *Heliyon*, 10(12), e32967. <https://doi.org/10.1016/j.heliyon.2024.e32967>
- Yang, Y. (2025). Sentiment analysis of consumer reviews on online shopping platforms using integrated deep learning models. *ICT Express*, 11(5), 881–887. <https://doi.org/10.1016/j.icte.2025.06.017>
- Zaid, S., Alharbi, A. H., & Samra, H. (2025). Multi-aspect sentiment classification of Arabic tourism reviews using BERT and classical machine learning. *Data*, 10(11), 168. <https://doi.org/10.3390/data10110168>