Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

Aspect-Based Sentiment Analysis Ulasan VGA RTX 5060 Berbasis LLM dan Naive Bayes

Ikhsan Fahruliazmi*[1], Dedi Saputra[2], Siti Nurdiani[3] , Juniato Sidauruk[4]
[1,2,3,4]Program Studi Informatika, Universitas Bina Sarana Informatika, Pontianak, Kalimantan Barat, Indonesia
*e-mail: 15220470@bsi.ac.id[1], dedi.dst@bsi.ac.id [2], siti.sxd@bsi.ac.id [3], juniato.jnd@bsi.ac.id [4]

Received:	Revised:	Accepted:	Available online:
07.06.2026	19.06.2026	25.06.2026	30.06.2026

Abstract: *Consumer reviews on e-commerce platforms contain valuable information about market perception of technology products. This study proposes an Aspect-Based Sentiment Analysis (ABSA) system for NVIDIA GeForce RTX 5060 GPU reviews from Tokopedia, Shopee, and YouTube using Complement Naive Bayes (CNB) with TF-IDF bigram feature representation. Main contributions include: (1) a multi-platform data collection pipeline yielding 4,675 original reviews, expanded to 4,982 reviews through data augmentation; (2) label quality improvement using Llama 3.1 LLM via Groq API achieving a 79.9% agreement rate (Cohen's Kappa = 0.699) with the keyword-based labels, with 20.1% of labels differing between the two methods; (3) Easy Data Augmentation (EDA) for balanced class distribution (Positive 33.5%; Neutral 32.7%; Negative 33.8%); and (4) identification of five review aspects: Price, Product Quality, Performance, Delivery, and Store Service. The global CNB model achieves 77.4% accuracy with 77.3% weighted F1-Score. Average ABSA accuracy per aspect reaches 91.9% with 92.5% average F1-Score, with highest performance on Price (95.4%) and lowest on Performance (86.7%). These results demonstrate the effectiveness of LLM integration in data labeling to enhance Naive Bayes-based ABSA systems.*
Keywords : *Aspect-Based Sentiment Analysis; Complement Naive Bayes; Large Language Model; TF-IDF; Easy Data Augmentation; VGA RTX 5060.*

Abstrak: Ulasan konsumen di platform e-commerce mengandung informasi berharga mengenai persepsi pasar terhadap produk teknologi. Penelitian ini mengusulkan sistem Aspect-Based Sentiment Analysis (ABSA) pada ulasan VGA NVIDIA GeForce RTX 5060 dari platform Tokopedia, Shopee, dan YouTube menggunakan metode Complement Naive Bayes (CNB) dengan representasi fitur TF-IDF bigram. Kontribusi utama meliputi: (1) *pipeline* pengumpulan data multi-platform yang menghasilkan 4.675 ulasan asli, diperkaya melalui augmentasi data menjadi 4.982 ulasan; (2) peningkatan kualitas label menggunakan Large Language Model Llama 3.1 melalui Groq API yang mencapai *agreement rate* 79,9% (Cohen's Kappa = 0,699), dengan 20,1% label berbeda dari pendekatan *keyword-based*; (3) augmentasi data menggunakan Easy Data Augmentation (EDA) untuk menyeimbangkan distribusi kelas menjadi hampir merata (Positif 33,5%; Netral 32,7%; Negatif 33,8%); serta (4) identifikasi lima aspek ulasan: Harga, Kualitas Produk, Performa, Pengiriman, dan Pelayanan Toko. Model CNB global mencapai akurasi 77,4% dengan F1-Score weighted 77,3%. Rata-rata akurasi ABSA per aspek mencapai 91,9% dengan rata-rata F1-Score 92,5%, dengan performa tertinggi pada aspek Harga (95,4%) dan terendah pada aspek Performa (86,7%). Hasil ini menunjukkan efektivitas integrasi LLM dalam proses pelabelan data untuk meningkatkan kualitas sistem ABSA berbasis Naive Bayes.
Kata kunci : Aspect-Based Sentiment Analysis; Complement Naive Bayes; *Large Language Model*; TF-IDF; Easy Data Augmentation; VGA RTX 5060.

1. PENDAHULUAN

Platform e-commerce di Indonesia mengalami pertumbuhan signifikan dengan jumlah pengguna internet mencapai 215 juta jiwa pada tahun 2023 (APJII, 2023). Ulasan konsumen merupakan sumber informasi kritis yang memengaruhi keputusan pembelian; penelitian menunjukkan produk dengan ulasan daring memiliki kemungkinan 270% lebih tinggi untuk dibeli (Chen et al., 2022). NVIDIA GeForce RTX 5060 merupakan kartu grafis berbasis arsitektur Blackwell yang diluncurkan tahun 2025 dengan 3.840 CUDA Cores, VRAM 8GB GDDR7, dan TDP 145W (NVIDIA Corporation, 2025), sehingga relevan untuk dianalisis sentimennya mengingat tingginya volume diskusi konsumen terhadap produk teknologi sejenis di platform e-commerce Indonesia.

Analisis sentimen konvensional mengklasifikasikan sentimen secara keseluruhan tanpa memisahkan aspek spesifik yang dibahas. Aspect-Based Sentiment Analysis (ABSA) mengatasi keterbatasan ini dengan mengidentifikasi aspek dan menentukan polaritas sentimen per aspek secara independen (Wankhade et al., 2022). Tantangan utama dalam membangun sistem ABSA Bahasa Indonesia meliputi keterbatasan data berlabel berkualitas tinggi, karakteristik bahasa informal, dan ketidakseimbangan distribusi kelas.

Penelitian ini mengusulkan *pipeline* yang mengintegrasikan Large Language Model (LLM) untuk meningkatkan kualitas pelabelan data secara signifikan. Model Llama 3.1-8b-instant melalui Groq API digunakan sebagai pengganti pendekatan *keyword-based* yang terbukti memiliki keterbatasan dalam menangkap nuansa sentimen. Complement Naive Bayes dipilih sebagai *classifier*

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

karena efisiensi komputasi, interpretabilitas tinggi, dan ketahanan terhadap ketidakseimbangan kelas (Dewi & Chen, 2022).

Kontribusi penelitian ini mencakup: (1) dataset ulasan VGA RTX 5060 berlabel LLM dari tiga platform dengan distribusi kelas seimbang; (2) evaluasi perbandingan kualitas label *keyword-based* versus LLM menggunakan Cohen's Kappa; (3) penerapan EDA untuk penyeimbangan kelas secara stratifikasi; serta (4) evaluasi ABSA komprehensif pada lima aspek ulasan produk GPU di Indonesia.

1.1 *Aspect-Based Sentiment Analysis*

ABSA merupakan pendekatan yang mengidentifikasi aspek dan menentukan polaritas sentimen per aspek dalam teks (Wankhade et al., 2022). Tinjauan analisis sentimen di era LLM (W. Zhang et al., 2023) mendefinisikan dua sub-tugas utama: identifikasi aspek dan klasifikasi sentimen per aspek. Berbeda dari sentiment analysis tingkat dokumen, ABSA memberikan informasi granular yang lebih bermanfaat bagi konsumen dan produsen. Penelitian ini menggunakan pendekatan berbasis kata kunci untuk identifikasi aspek dan Complement Naive Bayes untuk klasifikasi sentimen per aspek.

Tinjauan komprehensif oleh Poria et al. (2023) serta studi penerapan teknik klasifikasi oleh Danyal et al. (2023) meletakkan dasar metodologis bagi pengembangan teknik ABSA, termasuk ekstraksi fitur eksplisit dan penentuan orientasi sentimen pada tingkat kalimat. Brauwers dan Frasincar (Brauwers & Frasincar, 2023) lebih lanjut mengklasifikasikan pendekatan opinion mining menjadi tiga level analisis: dokumen, kalimat, dan aspek, dengan level aspek memberikan informasi paling granular bagi pengambilan keputusan bisnis.

1.2 Complement Naive Bayes dan TF-IDF

Naive Bayes merupakan algoritma klasifikasi probabilistik berbasis Teorema Bayes. Complement Naive Bayes (CNB) (Dewi & Chen, 2022) mengestimasi parameter dari data komplemen kelas, sehingga lebih robust terhadap ketidakseimbangan distribusi kelas. TF-IDF adalah metode pembobotan kata yang mempertimbangkan frekuensi kemunculan kata dalam dokumen dan kelangkaannya dalam koleksi (Xiao et al., 2024). Penelitian ini menggunakan kombinasi unigram-bigram dengan sublinear TF untuk merepresentasikan frasa bermakna seperti "performa bagus" sebagai fitur tunggal.

Kausar et al. (2023) pertama kali menunjukkan efektivitas Naive Bayes untuk klasifikasi sentimen dibandingkan pendekatan berbasis aturan. Pada domain ulasan film, Maulana et al. (2020) menunjukkan bahwa Support Vector Machine yang dikombinasikan dengan seleksi fitur Information Gain meningkatkan akurasi klasifikasi sentimen dibandingkan SVM tanpa seleksi fitur, mengindikasikan pentingnya pemilihan fitur yang relevan dalam klasifikasi teks. W. Wang et al. (2025) menunjukkan bahwa performa Naive Bayes pada data tidak seimbang dapat ditingkatkan melalui rekayasa fitur berbasis transformasi koordinat tanpa mengubah distribusi data asli, sebuah pendekatan yang relevan mengingat Complement Naive Bayes dirancang khusus untuk menangani ketidakseimbangan kelas pada dataset dengan ukuran *vocabulary* besar. Skema pembobotan TF-IDF yang digunakan dalam penelitian ini mengacu pada tinjauan information retrieval oleh Hambarde dan Proença (2023) yang menyeimbangkan frekuensi lokal dengan kelangkaan global suatu term.

1.3 *Large Language Model* untuk Pelabelan Data

Penggunaan LLM untuk anotasi data telah menunjukkan hasil yang menjanjikan sebagai alternatif anotasi manual yang mahal dan memakan waktu (S. Wang et al., 2023). LLM berbasis transformer mampu memahami konteks dan nuansa bahasa yang kompleks, termasuk sarkasme dan kritik halus yang sering terlewat oleh pendekatan berbasis kata kunci. Penelitian ini menggunakan Llama 3.1-8b-instant melalui Groq API dengan teknik *batch prompting* untuk efisiensi penggunaan kuota API.

Kemampuan LLM dalam memahami konteks bahasa berakar pada arsitektur Transformer yang ditinjau secara komprehensif oleh H. Zhang dan Shafiq (2024), yang menggantikan mekanisme rekurensi dengan *self-attention* untuk menangkap dependensi jarak jauh secara lebih efisien. Model BERT (Gardazi et al., 2025) memperluas arsitektur ini dengan *pre-training bidirectional*, sementara representasi vektor kata seperti word2vec (Touvron et al., 2023) menjadi fondasi awal bagi representasi semantik sebelum era Transformer. Untuk konteks Bahasa Indonesia, model IndoBERT (Ahmadian et al., 2024) menunjukkan performa kompetitif pada berbagai tugas NLP berbahasa Indonesia dan menjadi kandidat potensial untuk pengembangan lebih lanjut penelitian ini.

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

1.4 *Easy Data Augmentation*

Easy Data Augmentation (EDA), salah satu teknik augmentasi teks yang ditinjau dalam survei komprehensif oleh Bayer et al. (2022) mencakup empat operasi: penggantian sinonim, insersi acak, pertukaran acak, dan penghapusan acak. EDA terbukti meningkatkan performa model klasifikasi teks khususnya pada dataset kecil. Penelitian ini menerapkan EDA secara stratifikasi per kelas untuk mencapai distribusi yang seimbang tanpa kehilangan karakteristik linguistik data asli.

1.5 Penelitian Terkait

Beberapa penelitian telah menerapkan ABSA pada ulasan produk berbahasa Indonesia. Rahman et al. (2022) menerapkan ABSA pada ulasan produk e-commerce menggunakan metode SVM dan mencapai akurasi 82,51%. Bachtiar et al. (2022) menganalisis sentimen ulasan Free Fire menggunakan Naive Bayes Logistic dan melaporkan F1-Score 83,7%. Pangestu et al. (2023) menerapkan ABSA pada ulasan aplikasi dan menunjukkan pentingnya kualitas label dalam meningkatkan performa model. Penelitian ini memperbarui pendekatan tersebut dengan mengintegrasikan LLM untuk peningkatan kualitas label yang lebih sistematis.

2. METODE

2.1 Pengumpulan Data

Data dikumpulkan dari tiga platform menggunakan metode berbeda. Tokopedia (1.093 ulasan) menggunakan *web scraping* dengan Selenium dan undetected-chromedriver tanpa autentikasi. Shopee (238 ulasan) menggunakan teknik *anchor-based* dengan tombol "Laporkan Penyalahgunaan" sebagai penanda elemen ulasan asli. YouTube (3.344 komentar) menggunakan YouTube Data API v3 dengan empat lapis filter validasi video review RTX 5060 berbahasa Indonesia. Total data original yang terkumpul adalah 4.675 ulasan dari ketiga platform, sebelum augmentasi data.

2.2 *Preprocessing* Teks

Preprocessing dilakukan melalui enam tahap: (1) *case folding* yaitu konversi ke huruf kecil; (2) pembersihan yaitu menghapus URL, simbol, dan karakter non-alfabet; (3) normalisasi yaitu mengganti 87 pasang kata tidak baku menggunakan kamus yang dikurasi secara manual (contoh: "gak" → "tidak", "mantul" → "mantap"); (4) tokenisasi menggunakan NLTK word_tokenize; (5) penghapusan *stopword* menggunakan PySastrawi dengan 342 kata dasar ditambah 42 *stopword* domain-spesifik; (6) *stemming* menggunakan Sastrawi Stemmer.

Proses *stemming* Bahasa Indonesia menggunakan algoritma Confix Stripping yang ditinjau dan disempurnakan dalam kajian terbaru oleh Saputra et al. (2024), dengan penyempurnaan berdasarkan studi efektivitas stemming oleh Göksel et al. (2023) yang menunjukkan bahwa penerapan stemming secara selektif dapat meningkatkan recall pencarian informasi tanpa mengorbankan presisi. Tokenisasi dan pemrosesan teks lanjutan memanfaatkan pustaka NLTK (Mufti et al., 2026), sementara implementasi model klasifikasi menggunakan scikit-learn (Semary et al., 2024) yang menyediakan implementasi efisien untuk algoritma TF-IDF dan Complement Naive Bayes.

2.3 Pelabelan dengan *Large Language Model*

Pelabelan sentimen global dan per aspek dilakukan secara serentak menggunakan model llama-3.1-8b-instant melalui Groq API dengan teknik *batch prompting* (10 ulasan per *request*). Setiap ulasan diklasifikasikan ke Positif, Netral, atau Negatif untuk sentimen global, serta "Positif", "Netral", "Negatif", atau "N/A" untuk lima aspek (Harga, Kualitas Produk, Performa, Pengiriman, Pelayanan Toko). Kualitas pelabelan dievaluasi menggunakan Cohen's Kappa untuk mengukur tingkat kesepakatan antara metode *keyword-based* dan LLM.

Pemanggilan API menggunakan parameter *temperature* = 0 untuk meminimalkan variasi acak dalam output (deterministik) dan *max_tokens* = 1000 per request batch; parameter *top-p* tidak diatur secara eksplisit sehingga mengikuti nilai bawaan Groq API. Struktur prompt yang digunakan adalah sebagai berikut:

```
Kamu adalah sistem analisis sentimen Bahasa Indonesia untuk ulasan produk VGA (kartu grafis).
Analisis {n} ulasan berikut dan kembalikan hasil dalam format JSON array.

Untuk setiap ulasan tentukan:
- sentimen_global: "Positif", "Netral", atau "Negatif"
```

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

```
- Setiap aspek: "Positif", "Netral", "Negatif", atau "N/A" (tidak disebutkan)
  Aspek: Harga, Kualitas Produk, Performa, Pengiriman, Pelayanan Toko

Ulasan: {ulasan}

Jawab HANYA dengan JSON array (tanpa teks lain, tanpa markdown backtick):
[{"id":1,"sentimen_global":"...","Harga":"...","Kualitas Produk":"...","Performa":"...","Pengiriman":"...","Pelayanan Toko":"..."},...]
```

Variabel {n} dan {ulasan} disubstitusi secara dinamis dengan jumlah ulasan dalam batch (10) dan teks ulasan yang sedang diproses.

2.4 Augmentasi Data (EDA)

Augmentasi dilakukan secara stratifikasi per kelas menggunakan teknik EDA (Bayer et al., 2022) untuk mencapai distribusi kelas yang seimbang mendekati 5.000 data. Tiga operasi EDA diterapkan: (1) penggantian sinonim, yaitu mengganti kata dengan sinonimnya menggunakan kamus sinonim Bahasa Indonesia yang dikurasi; (2) penghapusan kata acak, yaitu menghapus 15% kata secara acak; (3) pertukaran kata acak, yaitu menukar posisi dua kata secara acak. Setiap data augmentasi mewarisi label sentimen dan aspek dari data sumber aslinya.

Pendekatan EDA dipilih dibandingkan teknik *oversampling* sintetis seperti SMOTE (Avela & Ilmonen, 2025) karena EDA bekerja langsung pada level teks tanpa memerlukan representasi vektor berdimensi tetap, sehingga lebih sesuai untuk data tekstual dengan panjang bervariasi.

2.5 Model Klasifikasi dan Evaluasi

Model utama menggunakan *pipeline* scikit-learn: TF-IDF (ngram_range=(1,2), max_features=5.000, sublinear_tf=True, min_df=2) → Complement Naive Bayes (alpha=0,3). Data dibagi dengan rasio 80:20 *stratified split*. Model ABSA dibangun terpisah untuk setiap aspek menggunakan subset data yang mengandung aspek tersebut. Evaluasi menggunakan akurasi, presisi, *recall*, dan F1-Score weighted average, dengan Stratified K-Fold (k=5) untuk validasi silang.

3. HASIL DAN PEMBAHASAN

3.1 Statistik Dataset

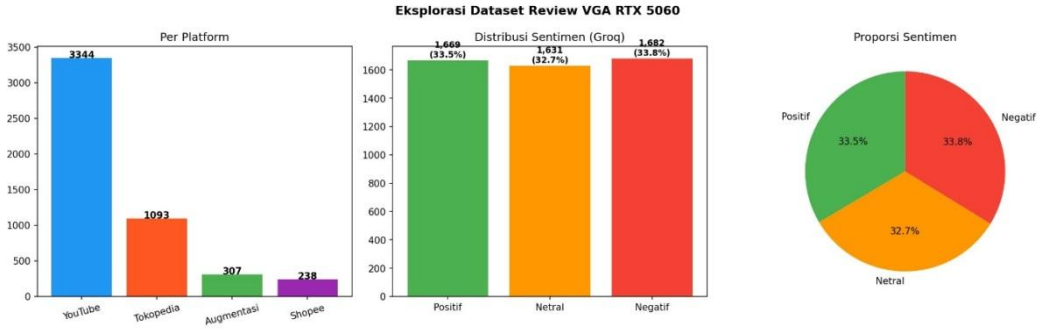
Dataset final terdiri dari 4.982 ulasan dari empat sumber data. Distribusi platform dan sentimen disajikan pada Tabel 1 dan Tabel 2.

Tabel 1. Distribusi Dataset per Sumber

Platform	Jumlah	Persentase (%)	Metode Pengumpulan
Tokopedia	1.093	21,9	Web Scraping (Selenium)
Shopee	238	4,8	Web Scraping (Anchor-Based)
YouTube	3.344	67,1	YouTube Data API v3
Augmentasi EDA	307	6,2	Easy Data Augmentation
Total	4.982	100	-

Tabel 2. Perbandingan Distribusi Sentimen Keyword vs LLM

Sentimen	Groq LLM	%	Keyword-Based	%
Positif	1.669	33,5	1.358	27,3
Netral	1.631	32,7	2.121	42,6
Negatif	1.682	33,8	1.503	30,2
Total	4.982	100	4.982	100



Gambar 1. Eksplorasi Dataset Review VGA RTX 5060 (Distribusi Platform dan Sentimen)

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

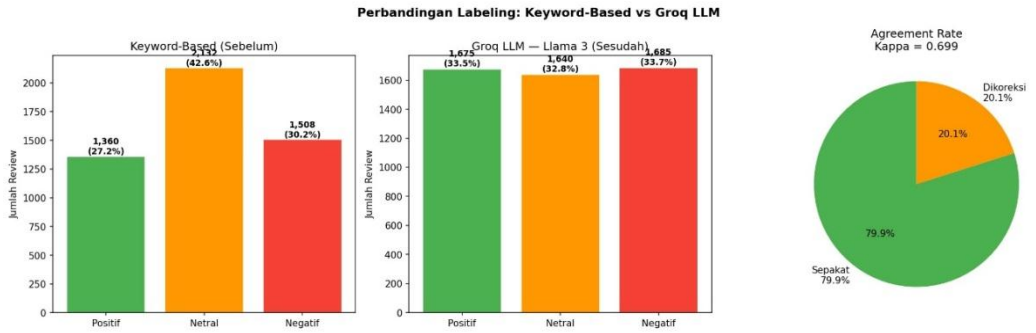
3.2 Perbandingan Kualitas Pelabelan

Perbandingan antara pelabelan *keyword-based* dan Groq LLM menunjukkan perbedaan distribusi yang signifikan. Metode *keyword* cenderung menghasilkan distribusi tidak seimbang dengan dominasi kelas Netral (42,6%), sedangkan LLM menghasilkan distribusi yang lebih merata. Cohen's Kappa sebesar 0,699 menunjukkan tingkat kesepakatan yang substansial antara kedua metode (Vach & Gerke, 2023), dengan LLM menghasilkan label berbeda pada 20,1% data *keyword-based*. Perbaikan paling signifikan terjadi pada pemisahan kelas Positif dan Negatif yang sebelumnya tumpang tindih pada kelas Netral oleh metode *keyword*.

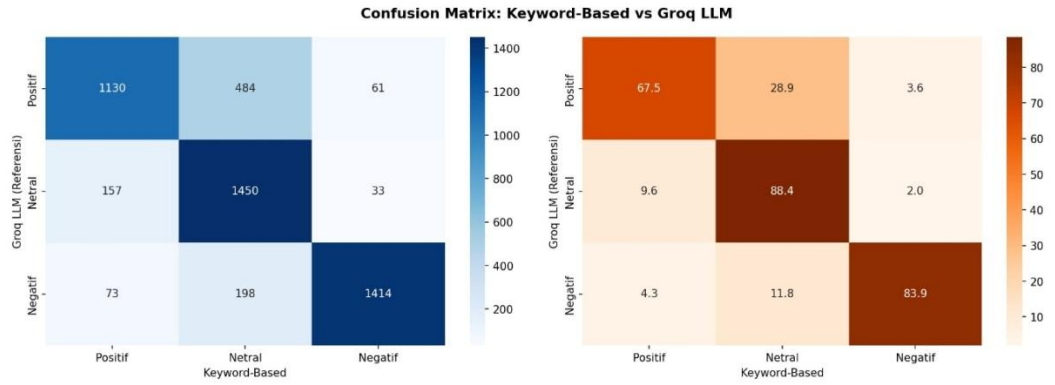
Untuk memberikan gambaran kualitatif terhadap 1.001 kasus perbedaan label, dilakukan inspeksi terhadap sampel acak kasus *disagreement*. Dua pola utama teridentifikasi: pertama, sekitar 10,6% kasus perbedaan melibatkan ulasan berbahasa Inggris (umumnya komentar YouTube), di mana metode *keyword-based* yang berbasis kamus kata kunci Bahasa Indonesia gagal mendeteksi sentimen eksplisit (misalnya "The best price per frame gpu on the market" diklasifikasikan Netral oleh keyword-based namun Positif oleh LLM). Kedua, sejumlah kasus menunjukkan kegagalan metode keyword-based dalam menangani negasi, seperti pada ulasan "Ga worth it udah tahun 2025 minimal 12 gb..." yang mengandung kata kunci positif ("worth it") namun didahului negasi ("Ga"), sehingga keyword-based salah mengklasifikasikan Positif sedangkan LLM benar mengklasifikasikan Negatif. Pola-pola ini memberikan indikasi kualitatif bahwa sebagian perbedaan label mencerminkan kelemahan metode keyword-based yang telah diketahui dalam literatur, bukan kesalahan acak. Meskipun demikian, inspeksi ini bersifat ilustratif dan dilakukan oleh peneliti, bukan oleh anotator independen; validasi yang lebih rigorus menggunakan beberapa anotator manusia independen dan pengukuran inter-rater reliability tetap diperlukan sebagai penelitian lanjutan untuk memastikan superioritas label LLM secara definitif.

Interpretasi nilai Kappa mengacu pada skala interpretasi yang dibahas dalam kajian metrik evaluasi terkini oleh Rainio et al. (2024), di mana nilai 0,699 yang diperoleh penelitian ini berada pada kategori "substansial" (0,61–0,80), mengindikasikan tingkat kesepakatan yang baik meskipun belum mencapai kategori "hampir sempurna".

Dari confusion matrix perbandingan, 67,5% data yang diberi label Positif oleh LLM sesuai dengan hasil *keyword*, sedangkan 28,9% sebelumnya berlabel Netral. Untuk kelas Negatif, 83,9% sesuai antara kedua metode, menunjukkan *keyword* cukup andal untuk kelas Negatif namun kurang akurat untuk membedakan Positif dan Netral.



Gambar 2. Perbandingan Distribusi Sentimen dan Agreement Rate: Keyword-Based vs Groq LLM



Gambar 3. Confusion Matrix Perbandingan Metode Keyword-Based vs Groq LLM

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

3.3 Evaluasi Model Complement Naive Bayes Global

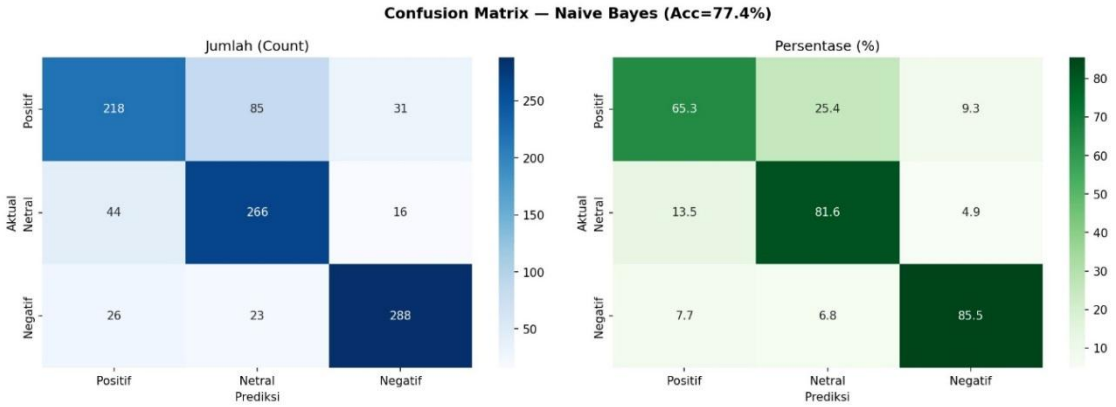
Tabel 3 menyajikan hasil evaluasi model CNB untuk klasifikasi sentimen global. Model mencapai akurasi 77,4% dengan F1-Score weighted 77,3%.

Tabel 3. Evaluasi Model Complement Naive Bayes Global

Kelas Sentimen	Presisi (%)	Recall (%)	F1-Score (%)	Support
Positif	75,69	65,27	70,10	334
Netral	71,12	81,60	76,00	326
Negatif	85,97	85,46	85,71	337
Weighted Avg	77,67	77,43	77,31	997
Accuracy	-	-	77,40	997

Kelas Negatif mencapai F1-Score tertinggi (85,71%) berkat distribusi data yang seimbang hasil augmentasi EDA dan kualitas label Groq yang akurat. Kelas Positif memiliki F1-Score terendah (70,10%) karena terdapat tumpang tindih semantik dengan kelas Netral, khususnya pada ulasan yang bersifat deskriptif tanpa evaluasi eksplisit. Kelas Netral menunjukkan recall tinggi (81,60%) namun presisi lebih rendah (71,12%), mengindikasikan model cenderung mengklasifikasikan ulasan ambigu ke kelas Netral.

Metrik evaluasi yang digunakan, yaitu presisi, recall, dan F1-Score, mengikuti kerangka kerja evaluasi klasifikasi yang dijelaskan Cabot dan Ross (Cabot & Ross, 2023), di mana F1-Score sebagai rata-rata harmonik presisi dan recall memberikan gambaran tunggal yang menyeimbangkan kedua metrik tersebut, sangat relevan untuk dataset dengan distribusi kelas yang telah diseimbangkan.



Gambar 4. Confusion Matrix Model Complement Naive Bayes (Akurasi Global 77,4%)

3.4 Hasil ABSA Per Aspek

Model ABSA dibangun terpisah untuk setiap dari lima aspek yang diidentifikasi menggunakan kata kunci domain-spesifik. Tabel 4 menyajikan hasil evaluasi model per aspek beserta distribusi data.

Tabel 4. Hasil Evaluasi ABSA Per Aspek

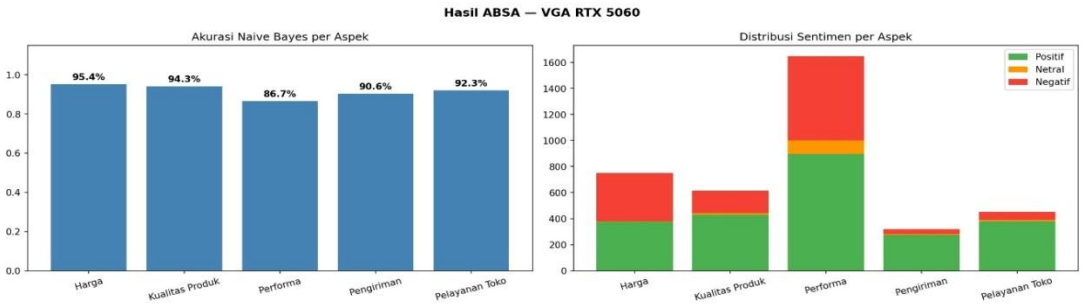
Aspek	N Data	Positif	Netral	Negatif	Acc (%)	Prec (%)	Recall (%)	F1 (%)
Harga	752	376	4	372	95,36	95,44	95,36	95,36
Kualitas Produk	613	431	11	171	94,31	92,77	94,31	93,51
Performa	1.648	898	102	648	86,67	87,08	86,67	86,39
Pengiriman	319	274	8	37	90,60	96,40	90,60	92,58
Pelayanan Toko	454	381	7	66	92,31	98,29	92,31	94,57
Rata-rata	-	-	-	-	91,85	94,00	91,85	92,48

Aspek Harga mencapai akurasi tertinggi (95,4%) dengan distribusi data yang nyaris biner antara Positif dan Negatif (hanya 4 data Netral dari 752), sehingga model lebih mudah melakukan klasifikasi. Aspek Pengiriman mencapai presisi tertinggi (96,4%) menunjukkan model sangat akurat dalam memprediksi sentimen pengiriman meskipun datanya terbatas (319 ulasan). Aspek Performa memiliki akurasi terendah (86,7%) akibat ambiguitas tinggi: ulasan performa sering bersifat kontekstual dan

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

bergantung pada kebutuhan pengguna (gaming 1080p vs 4K), serta adanya 102 data Netral yang sulit dibedakan dari Positif.

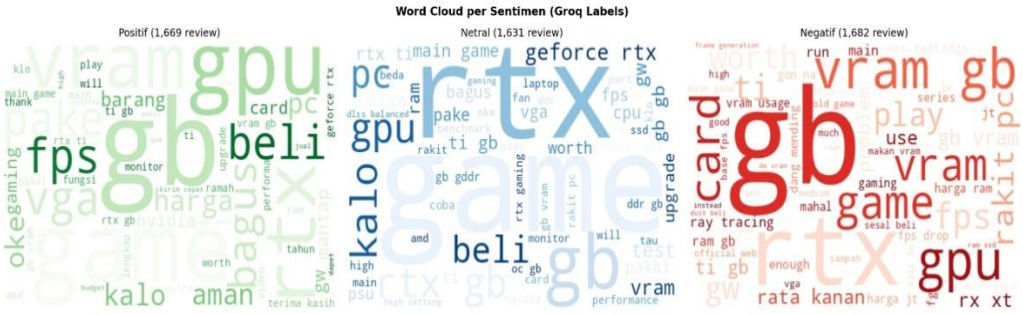
Secara keseluruhan, rata-rata akurasi ABSA sebesar 91,9% dan F1-Score 92,5% jauh melampaui akurasi klasifikasi global (77,4%), mengkonfirmasi bahwa pemisahan per aspek meningkatkan kemampuan klasifikasi model secara signifikan (Wankhade et al., 2022). Aspek Performa juga memiliki data terbanyak (1.648 ulasan, 33,1% dataset) mengindikasikan bahwa performa GPU merupakan aspek yang paling banyak dibahas konsumen Indonesia dalam ulasan VGA RTX 5060.



Gambar 5. Akurasi Naive Bayes dan Distribusi Sentimen per Aspek

3.5 Analisis *Word Cloud*

Analisis word cloud per kelas sentimen mengungkapkan pola linguistik yang berbeda. Pada sentimen Positif (1.669 ulasan), kata dominan adalah "gpu", "fps", "beli", "vram", dan "gb", menunjukkan konsumen merasa puas dengan performa gaming dan keputusan pembelian. Pada sentimen Netral (1.631 ulasan), kata dominan "rtx", "kalo", "game", dan "worth" mengindikasikan ulasan yang bersifat komparatif dan pertanyaan kelayakan pembelian. Pada sentimen Negatif (1.682 ulasan), kata "vram", "gb", "mahal", dan "card" mendominasi, mencerminkan kekhawatiran konsumen terhadap kapasitas VRAM 8GB yang dianggap kurang memadai untuk gaming modern dan persepsi harga yang tidak sebanding dengan peningkatan performa dibandingkan generasi sebelumnya.



Gambar 6. Word Cloud per Kelas Sentimen (Groq Labels): Positif, Netral, Negatif

3.6 Validasi Lintas Domain dan Perbandingan Produk (RTX 4060)

Validasi lintas domain bertujuan mengukur kemampuan generalisasi model ABSA yang dilatih pada data RTX 5060 ketika diterapkan pada data ulasan produk GPU generasi sebelumnya, yaitu NVIDIA GeForce RTX 4060. Kedua produk berada pada segmen pasar yang sama (kelas menengah) namun berbeda generasi arsitektur, sehingga perbandingan ini relevan untuk menilai stabilitas performa model di luar domain data pelatihan dan untuk memahami perbedaan persepsi konsumen terhadap kedua generasi produk.

3.6.1 Hasil Validasi Lintas Domain

Pengujian *cross-domain* dilakukan dengan menerapkan model Complement Naive Bayes yang telah dilatih pada dataset RTX 5060 (3.985 data *training* dari 4.982 data, *split* 80:20) untuk mengklasifikasikan ulasan RTX 4060 (2.814 data) yang dikumpulkan secara independen dari platform Tokopedia dan YouTube. Dataset RTX 4060 diberi label menggunakan metode yang identik dengan

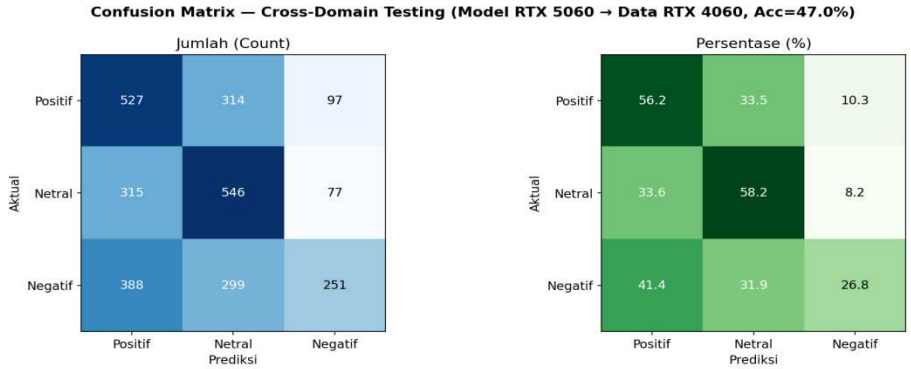
Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

RTX 5060, yaitu Groq LLM dengan model llama-3.1-8b-instant, untuk memastikan konsistensi evaluasi.

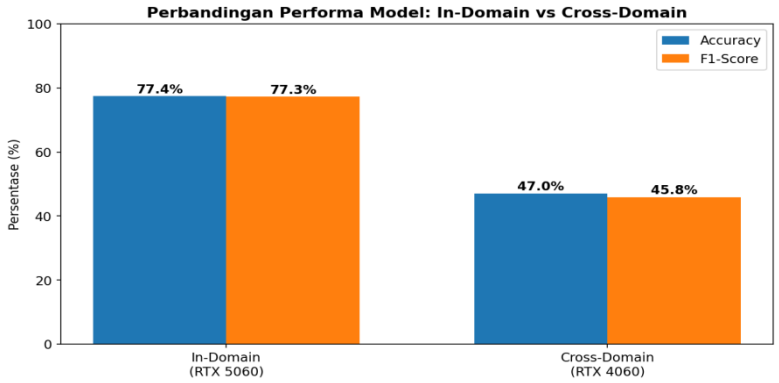
Tabel 5. Perbandingan Akurasi Model: In-Domain vs Cross-Domain

Evaluasi	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
In-Domain (RTX 5060)	77,43	77,61	77,43	77,28
Cross-Domain (RTX 4060)	47,05	49,67	47,05	45,84

Hasil pengujian *cross-domain* pada Tabel 5 menunjukkan penurunan akurasi yang signifikan dari 77,43% (*in-domain*) menjadi 47,05% (*cross-domain*), atau penurunan sebesar 30,38 poin persentase. Penurunan F1-Score juga serupa, dari 77,28% menjadi 45,84% (selisih 31,44 poin). Hasil ini mengindikasikan bahwa model global mengalami *overfitting* yang substansial terhadap karakteristik linguistik spesifik ulasan RTX 5060, sehingga kemampuan generalisasinya terhadap produk lain dalam kategori serupa masih terbatas. Penurunan ini kemungkinan disebabkan oleh perbedaan *vocabulary* spesifik produk (misalnya istilah yang merujuk pada arsitektur atau fitur khas RTX 5060) yang tidak ditemukan pada ulasan RTX 4060.



Gambar 7. Confusion Matrix Cross-Domain: Model RTX 5060 Diuji pada Data RTX 4060



Gambar 8. Perbandingan Performa Model: In-Domain vs Cross-Domain

3.6.2 Perbandingan Sentimen per Aspek RTX 5060 vs RTX 4060

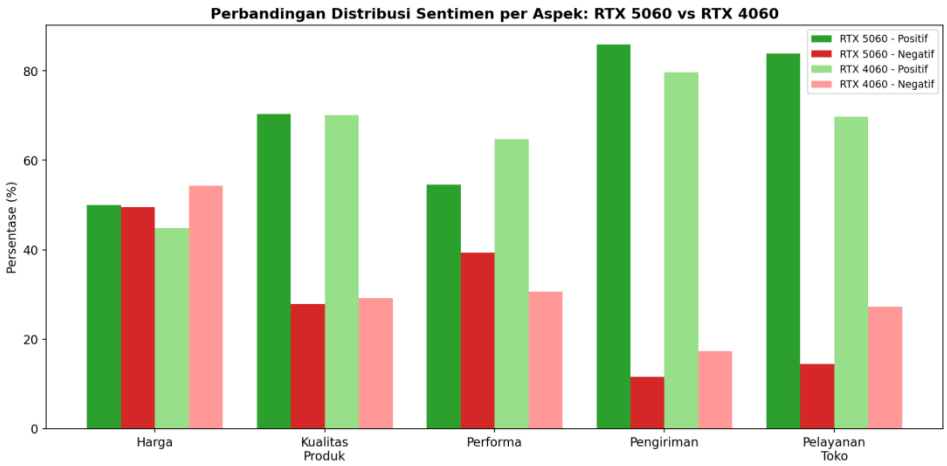
Tabel 6. Perbandingan Sentimen per Aspek: RTX 5060 vs RTX 4060

Aspek	RTX 5060 Positif (%)	RTX 5060 Negatif (%)	RTX 4060 Positif (%)	RTX 4060 Negatif (%)
Harga	50,0	49,5	44,9	54,3
Kualitas Produk	70,3	27,9	70,1	29,2
Performa	54,5	39,3	64,7	30,6
Pengiriman	85,9	11,6	79,7	17,3
Pelayanan Toko	83,9	14,5	69,7	27,2

Tabel 6 menyajikan perbandingan distribusi sentimen per aspek antara kedua produk. Aspek Harga menunjukkan perbedaan paling mencolok: RTX 5060 memiliki sentimen Negatif 49,5% sedangkan RTX 4060 lebih tinggi (54,3%), mengindikasikan konsumen menilai kedua produk relatif

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

kurang terjangkau, dengan RTX 4060 sedikit lebih banyak dikritik dari aspek harga meskipun merupakan produk generasi lebih lama. Sebaliknya, aspek Performa pada RTX 4060 menunjukkan sentimen Positif lebih tinggi (64,7% berbanding 54,5% pada RTX 5060), kemungkinan karena ekspektasi konsumen terhadap RTX 4060 yang lebih moderat dibandingkan RTX 5060 yang dipasarkan sebagai produk generasi terbaru. Aspek Pelayanan Toko pada RTX 5060 (83,9% Positif) jauh lebih baik dibandingkan RTX 4060 (69,7% Positif), yang dapat dipengaruhi oleh perbedaan reputasi penjual pada data yang dikumpulkan untuk masing-masing produk.



Gambar 9. Perbandingan Distribusi Sentimen per Aspek: RTX 5060 vs RTX 4060

Tabel 7. Hasil Evaluasi ABSA Per Aspek: In-Domain vs Cross-Domain

Aspek	N (4060)	Acc In (%)	Acc Cross (%)	Prec Cross (%)	Rec Cross (%)	F1 Cross (%)
Harga	613	96,69	53,02	55,01	53,02	51,89
Kualitas Produk	432	93,50	68,06	60,19	68,06	60,50
Performa	731	87,88	63,89	59,80	63,89	57,88
Pengiriman	197	92,19	75,63	72,45	75,63	73,81
Pelayanan Toko	254	95,60	72,44	68,43	72,44	68,61
Rata-rata	-	93,17	66,61	63,18	66,61	62,54

Evaluasi ABSA per aspek pada skema *cross-domain* (Tabel 7) menunjukkan variasi penurunan performa yang berbeda antar aspek. Aspek Pengiriman mempertahankan performa terbaik dengan akurasi *cross-domain* 75,63% (F1=73,81%), hanya turun 16,55 poin dari *in-domain* (92,19%), mengindikasikan bahwa kosakata terkait pengiriman bersifat umum dan tidak spesifik produk. Sebaliknya, aspek Harga mengalami penurunan paling signifikan dari 96,69% menjadi 53,02% (turun 43,67 poin), mencerminkan bahwa ekspresi terkait harga sangat bergantung pada konteks harga pasar aktual masing-masing produk, sehingga model yang dilatih pada rentang harga RTX 5060 kesulitan menggeneralisasi pada rentang harga RTX 4060 yang berbeda. Rata-rata akurasi ABSA *cross-domain across* lima aspek adalah 66,61% dengan F1-Score 62,54%, masih lebih baik dibandingkan akurasi sentimen global *cross-domain* (47,05%), mengonfirmasi bahwa pemisahan klasifikasi per aspek tetap memberikan keunggulan meskipun pada skenario lintas domain.

3.7 Validasi Metodologis: Pengaruh Urutan Augmentasi terhadap Hasil

Pipeline utama pada Sub-bab 2.4-2.5 menerapkan augmentasi EDA sebelum pembagian data latih-uji. Urutan ini berpotensi menimbulkan kebocoran data (*data leakage*) apabila data hasil augmentasi dan data sumber aslinya terpisah ke set latih dan uji yang berbeda, sehingga model berisiko menghasilkan estimasi performa yang terlalu optimis. Untuk memvalidasi hal ini, dilakukan eksperimen ulang dengan urutan yang dikoreksi: pembagian data 80:20 dilakukan terlebih dahulu pada 4.675 data asli (sebelum augmentasi), kemudian augmentasi EDA hanya diterapkan pada subset data latih (kelas Negatif yang kurang representatif) untuk mencapai keseimbangan kelas, sedangkan data uji (935 data) tidak pernah tersentuh proses augmentasi maupun proses pelatihan model. Hasil

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

eksperimen dengan urutan terkoreksi ini menghasilkan akurasi 78,40% dengan F1-Score weighted 78,19%, dibandingkan 77,40% dan 77,31% pada eksperimen utama dengan urutan augmentasi-sebelum-split. Selisih sebesar +1,00 poin persentase berada dalam rentang variasi normal akibat perbedaan komposisi set uji, dan tidak menunjukkan penurunan performa yang mengindikasikan kebocoran data signifikan. Hasil ini mengindikasikan bahwa operasi augmentasi EDA yang digunakan (penghapusan kata acak dan pertukaran kata acak) menghasilkan variasi teks yang cukup berbeda dari data sumber sehingga tidak menyebabkan duplikasi mendekati identik antara set latih dan uji. Meskipun demikian, untuk praktik terbaik pada penelitian selanjutnya, urutan pembagian data sebelum augmentasi tetap direkomendasikan sebagai pendekatan yang lebih konservatif dan metodologis lebih dapat dipertanggungjawabkan.

4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan sistem Aspect-Based Sentiment Analysis pada ulasan VGA NVIDIA GeForce RTX 5060 menggunakan Complement Naive Bayes dengan peningkatan kualitas label berbasis Large Language Model. Beberapa temuan utama yang dapat disimpulkan:

Pertama, integrasi LLM (Llama 3.1 via Groq API) dalam proses pelabelan menghasilkan *agreement rate* 79,9% (Cohen's Kappa = 0,699) dengan metode *keyword-based*, dengan 20,1% label berbeda dari pendekatan keyword-based dan distribusi kelas yang lebih representatif. Temuan ini mengindikasikan LLM berpotensi menangkap nuansa sentimen yang tidak tertangkap oleh pendekatan berbasis kata kunci, meskipun validasi lebih lanjut terhadap anotasi manusia diperlukan untuk memastikan keunggulan ini secara definitif.

Kedua, model CNB global mencapai akurasi 77,4% dengan F1-Score weighted 77,3%, dengan performa terbaik pada kelas Negatif (F1=85,7%) yang mendapat manfaat terbesar dari penyeimbangan kelas melalui augmentasi EDA.

Ketiga, rata-rata akurasi ABSA per aspek sebesar 91,9% melampaui akurasi global secara signifikan, dengan rentang dari 86,7% (Performa) hingga 95,4% (Harga). Aspek Performa merupakan topik paling banyak dibahas konsumen (33,1% data) dengan dominasi kekhawatiran terhadap VRAM 8GB.

Keempat, validasi lintas domain pada data RTX 4060 mengungkapkan penurunan performa signifikan (akurasi global dari 77,43% menjadi 47,05%), mengindikasikan model mengalami *overfitting* terhadap karakteristik linguistik spesifik RTX 5060. Namun demikian, pemisahan klasifikasi per aspek tetap mempertahankan keunggulan relatif pada skenario *cross-domain* (rata-rata akurasi ABSA 66,61% dibandingkan akurasi global 47,05%), dengan aspek Pengiriman menunjukkan generalisasi terbaik (75,63%) dan aspek Harga menunjukkan generalisasi terlemah (53,02%) akibat ketergantungan tinggi terhadap konteks harga pasar spesifik produk.

Penelitian selanjutnya disarankan menerapkan model berbasis IndoBERT atau transformer untuk Bahasa Indonesia, memperluas jumlah aspek, serta menambahkan validasi lintas domain menggunakan data ulasan produk GPU generasi sebelumnya. Algoritma klasifikasi berbasis pohon keputusan seperti Random Forest dan C4.5 yang terbukti efektif pada domain data terstruktur lain (Linawati et al., 2020) juga dapat dieksplorasi sebagai pembanding performa terhadap Complement Naive Bayes pada penelitian mendatang.

DAFTAR PUSTAKA

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language. *Applied Computational Intelligence and Soft Computing*, *2024*, 2826773. https://doi.org/10.1155/2024/2826773

APJII. (2023). *Survei Penetrasi Internet Indonesia 2022-2023*. https://apjii.or.id/survei

Avela, A., & Ilmonen, P. (2025). Extrapolated Markov Chain Oversampling Method for Imbalanced Text Classification. *ArXiv Preprint ArXiv:2509.02332*.

Bachtiar, A., & others. (2022). Analisis Sentimen Ulasan Free Fire Menggunakan Naive Bayes Logistic. *Prosiding Seminar Nasional Teknologi Dan Sistem Informasi (Semnasteknomedia)*.

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, *55*(7), 146. https://doi.org/10.1145/3544558

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

Brauwers, G., & Frasincar, F. (2023). A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, *35*(3), 2461–2482. https://doi.org/10.1109/TKDE.2022.3230975

Cabot, J. H., & Ross, E. G. (2023). Evaluating Prediction Model Performance. *Surgery*, *174*(3), 723–726. https://doi.org/10.1016/j.surg.2023.05.023

Chen, T., Samaranayake, P., Cen, X., Qi, M., & Lan, Y.-C. (2022). The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence From an Eye-Tracking Study. *Frontiers in Psychology*, *13*, 865702. https://doi.org/10.3389/fpsyg.2022.865702

Danyal, M. M., Khan, S. S., Khan, M., Ghaffar, M. B., Khan, B., & Arshad, M. (2023). Sentiment Analysis Based on Performance of Linear Support Vector Machine and Multinomial Naïve Bayes Using Movie Reviews with Baseline Techniques. *Journal of Big Data*, *5*(1), 1–18. https://doi.org/10.32604/jbd.2023.041319

Dewi, C., & Chen, R.-C. (2022). Complement Naive Bayes Classifier for Sentiment Analysis of Internet Movie Database. *Intelligent Information and Database Systems (ACIIDS 2022), Lecture Notes in Computer Science*, *13757*, 81–93. https://doi.org/10.1007/978-3-031-21743-2_7

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: a review. *Artificial Intelligence Review*, *58*(6), 166. https://doi.org/10.1007/s10462-025-11162-5

Göksel, G., Arslan, A., & Dinçer, B. T. (2023). A selective approach to stemming for minimizing the risk of failure in information retrieval systems. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.1175

Hambarde, K. A., & Proença, H. (2023). Information Retrieval: Recent Advances and Beyond. *IEEE Access*, *11*, 76581–76604. https://doi.org/10.1109/ACCESS.2023.3295776

Kausar, M. A., Fageeri, S. O., & Soosaimanickam, A. (2023). Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews. *Engineering, Technology \& Applied Science Research*, *13*(3), 10849–10855. https://doi.org/10.48084/etasr.5854

Linawati, S., Nurdiani, S., Handayani, K., & Latifah, L. (2020). Prediksi Prestasi Akademik Mahasiswa Menggunakan Algoritma Random Forest Dan C4.5. *Jurnal Khatulistiwa Informatika*, *8*(1), 47–52. https://doi.org/10.31294/jki.v8i1.7827

Maulana, R., Rahayuningsih, P. A., Irmayani, W., Saputra, D., & Jayanti, W. E. (2020). Improved Accuracy of Sentiment Analysis Movie Review Using Support Vector Machine Based Information Gain. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012060

Mufti, M., Hammad, O., & Rahman, M. (2026). Improving User Experience with Personalized Review Ranking and Summarization. *ArXiv Preprint ArXiv:2601.05261*.

NVIDIA Corporation. (2025). NVIDIA GeForce RTX 5060 Series – Blackwell Architecture. In *NVIDIA Technical Documentation*. https://www.nvidia.com/en-us/geforce/graphics-cards/50-series/

Pangestu, A. S., Christian, E., & Lestari, A. (2023). Aspect-Based Sentiment Analysis pada Ulasan Aplikasi. *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *5*(1), 45–52.

Poria, S., Hazarika, D., Majumder, N., & Mihalcea, R. (2023). Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research. *IEEE Transactions on Affective Computing*, *14*(1), 108–132. https://doi.org/10.1109/TAFFC.2020.3038167

Rahman, R. A., Pranatawijaya, V. H., & Sari, N. N. K. (2022). Analisis Sentimen Berbasis Aspek pada Ulasan Produk E-Commerce. *Jurnal Informatika*, *18*(2). https://doi.org/10.26555/jifo.v18i2

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Saputra, W. A. M., Utami, E., & Yaqin, A. (2024). Unlocking Insights: A Literature Review on Enhanced Confix Stripping and Nazief \& Adriani Algorithm Modifications for Makassar Language Text Stemming. *International Journal of Innovative Science and Research Technology*, 603–610.

Semary, N. A., Ahmed, W., Amin, K., & Pławiak Paweł and Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. *PLOS ONE*, *19*(2), e0294968. https://doi.org/10.1371/journal.pone.0294968

Vol. 6, No. 1 - Juni 2026, Hal. 65-76
DOI: 10.59395/jitp.v6i1.198

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*.

Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa -- A comparison of basic properties. *MethodsX*, *10*, 102212. https://doi.org/10.1016/j.mex.2023.102212

Wang, S., Liu, Z., Zheng, W., Gan, Z., Wei, F., & Tang, J. (2023). *Is ChatGPT a Good Sentiment Reasoner? A Preliminary Study*. https://arxiv.org/abs/2304.04339

Wang, W., Yan, L., Liu, F., & Li, Y. (2025). Improving Gaussian Naive Bayes classification on imbalanced data through coordinate-based minority feature mining. *PeerJ Computer Science*, *11*, e3003. https://doi.org/10.7717/peerj-cs.3003

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A Survey on Sentiment Analysis Methods, Applications, and Challenges. *Artificial Intelligence Review*, *55*, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Xiao, L., Li, Q., Ma, Q., Shen, J., Yang, Y., & Li, D. (2024). Text classification algorithm of tourist attractions subcategories with modified TF-IDF and Word2Vec. *PLoS ONE*, *19*(10), e0305095. https://doi.org/10.1371/journal.pone.0305095

Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*, *11*(1), 25. https://doi.org/10.1186/s40537-023-00842-0

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment Analysis in the Era of Large Language Models: A Reality Check. *ArXiv Preprint*.