Perbandingan Performa LLaMA-2 dan GPT-3.5 Turbo Menggunakan Metode *Retrieval Augmented Few-shot* pada Analisis Sentimen

I Wayan Adi Maha Wiguna [a,1,*], Ida Bagus Nyoman Pascima[b,2], Luh Putu Eka Damayanti[c,3]

[a,b,c] Pendidikan Teknik Informatika, Fakultas Teknik dan Kejuruan, Universitas Pendidikan Ganesha
[1] adi.maha@student.undiksha.ac.id*; [2] gus.pascima@undiksha.ac.id; [3] ekadamayanthi@undiksha.ac.id
* Penulis Korespondensi

ABSTRAK

Large Language Models (LLM) memerlukan metode tambahan untuk optimasi pada tugas spesifik seperti analisis sentimen. Penelitian ini membandingkan performa GPT-3.5 Turbo dan LLaMA-2 melalui penerapan metode *Retrieval Augmented Few-shot* (RAFS) pada domain pariwisata, dengan skenario Zero-shot sebagai baseline. Hasil eksperimen menunjukkan bahwa LLaMA-2 mengalami peningkatan performa yang jauh lebih signifikan dibandingkan GPT-3.5 Turbo setelah penerapan RAFS. Akurasi LLaMA-2 meningkat dari 0,833 menjadi 0,862, sementara GPT-3.5 Turbo hanya meningkat tipis dari 0,851 menjadi 0,856. Perbedaan substansial terlihat pada metrik kelas minoritas; *f1-score* GPT-3.5 hanya naik dari 0,555 ke 0,572, sedangkan LLaMA-2 melonjak drastis dari 0,462 ke 0,676 dengan kenaikan presisi dari 0,395 ke 0,844. Secara *head-to-head*, LLaMA-2 terbukti sedikit lebih unggul dibanding dengan GPT-3.5 Turbo dalam menghasilkan klasifikasi yang tepat dan seimbang. Meskipun GPT-3.5 memiliki baseline awal yang lebih tinggi, LLaMA-2 menunjukkan kemampuan adaptasi dan skalabilitas yang lebih baik terhadap augmentasi konteks. Temuan ini menegaskan bahwa model *open-source* dengan dukungan RAFS mampu menyamai, bahkan melampaui model *proprieter* dalam menangani kompleksitas sentimen ulasan pelanggan.

Riwayat Artikel
Diterima 25 Januari 2026
Diperbaiki 15 Februari 2026
Diterbitkan 25 Februari 2026

Kata Kunci
LLM
Analisis Sentimen
GPT
LLaMA
Retrieval
Few-shot

1. Pendahuluan

Pada era digital jutaan ulasan, komentar dan opini telah dibagikan oleh pengguna melalui platform internet. Data yang tersebar tersebut memiliki nilai strategis jika dimanfaatkan dengan baik, salah satunya adalah ulasan pelanggan. Ulasan pelanggan merupakan salah satu tolak ukur dalam mengambil keputusan strategis khususnya pada industri akomodasi. Untuk mengetahui bagaimana emosi yang terkandung dari sebuah ulasan dapat dilakukan dengan menganalisis sentimen ulasan tersebut.

Analisis sentimen adalah bidang studi yang berfokus menganalisis opini, sentimen, emosi, penilaian, dan sikap seseorang terhadap suatu topik [1] [2]. Analisis sentimen telah banyak digunakan oleh perusahaan untuk mengetahui review dari produk atau jasa yang ditawarkan [3]. Analisis sentimen dapat dilakukan dengan memanfaatkan machine learning tradisional seperti *Support Vector Machine*, *Naive Bayes*, atau *Random Forest*. Namun, menggunakan metode tradisional untuk pengolahan data teks menghadapi beberapa tantangan, terutama dalam hal akurasi dan pemahaman berbagai bahasa [4]. Permasalahan tersebut dapat diatasi dengan menggunakan model besar yang sudah dilatih dengan banyak data sebelumnya, model tersebut bernama *Large Language Model* (LLM). LLM adalah model yang dilatih dengan kumpulan data yang besar, LLM juga dapat memahami bahasa alami atau NLP dengan tingkat pemahaman bahasa dan pengetahuan yang luas [5].

Dalam tugas analisis sentimen LLM dapat dijalankan dengan berbagai metode, seperti *Zero-shot*, *Fine Tuning*, *Few-shot* atau *Retrieval Augmented Generation* (RAG). Metode seperti *Zero-shot* walau mudah untuk diterapkan tanpa data latih namun memiliki performa yang kurang baik sedangkan penerapan *Few-shot* dalam kasus analisis sentimen memerlukan komputasi yang tinggi karena memproses data keseluruhan contoh dan data ujinya tanpa memikirkan kesamaan semantik data. Disisi lain *Fine-tuning* dapat memberikan performa yang lebih baik namun diperlukannya komputasi yang besar serta dataset yang sangat banyak. Sedangkan pendekatan RAG kurang efektif karena tidak terdapat contoh label pada data retrievalnya. Penelitian oleh [6] membuktikan bahwa performa RAG dapat ditingkatkan lebih lanjut lagi, terutama dengan memberikan penalaran berbasis contoh. Selain itu penelitian oleh [7] menunjukan bahwa hasil dari metode *Few-shot* sangat bergantung pada relevansi dari contoh yang dipilih. Kedua penelitian tersebut menunjukan bahwa metode RAG dan *Few-shot* dapat saling melengkapi untuk tugas analisis sentimen.

Oleh karena itu diperlukannya sebuah pendekatan yang lebih canggih yaitu *Retrieval Augmented Few-shot* (RAFS). Pendekatan RAFS ini secara dinamis mengambil contoh ulasan berlabel yang paling relevan secara semantik dari *knowledge base* untuk setiap data uji yang unik [8]. Pendekatan ini mengadopsi mekanisme *retrieval* dari RAG untuk mengekstraksi informasi yang paling relevan secara semantik, sementara teknik *Few-shot* digunakan untuk memandu penalaran model melalui penyertaan contoh ulasan dan label dari hasil *retrieval* secara eksplisit di dalam prompt. Sistem serupa telah dikembangkan oleh [9] didapat bahwa adanya peningkatan rata-rata yang paling signifikan yaitu pada model Gpt-neo-2.7b yang meningkat sebanyak 7.4 poin.

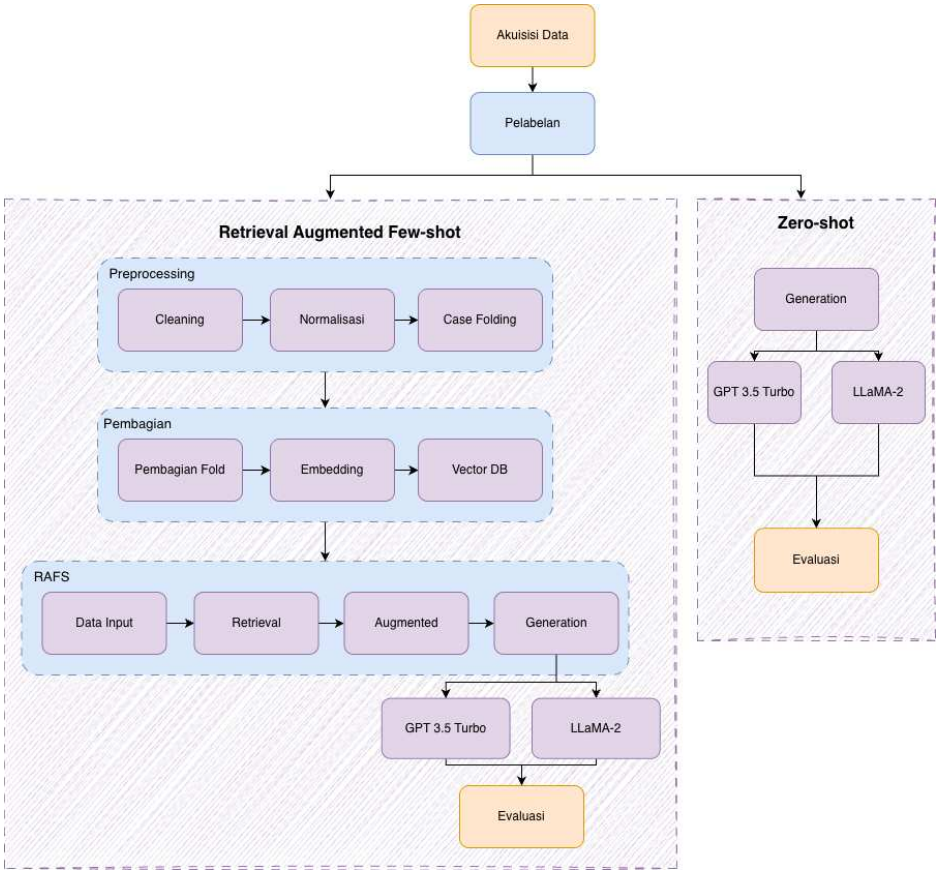
Pada penelitian ini model yang akan dibandingkan adalah LLaMA-2 dan GPT-3.5 Turbo. Hal ini kemudian memunculkan pertanyaan baru yang krusial, mengingat dua LLM terkemuka seperti LLaMA dan GPT dikembangkan oleh arsitektur dan data pelatihan yang berbeda oleh dua perusahaan yang berbeda pula, maka tidak ada jaminan kedua model tersebut dapat menghasilkan merespon yang sama saat menggunakan metode RAFS. Adanya celah penelitian berupa minimnya penelitian yang menyajikan perbandingan *head-to-head* yang secara empiris mengukur performa kedua LLM yakni LLaMA-2 dan GPT-3.5 Turbo dalam menerapkan metode RAFS yang terkontrol. Setidaknya ada 2 permasalahan yang akan dibahas pada penelitian ini (1) Apakah ada peningkatan yang signifikan setelah penerapan metode RAFS, (2) Model manakah yang menghasilkan performa yang lebih baik setelah penerapan metode RAFS. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi praktis yang berharga berupa benchmark berbasis data bagi para pengembang dan praktisi industri dalam membuat keputusan strategis saat memilih platform LLM untuk aplikasi sejenis.

2. Metode

2.1 Alur Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimental yang berfokus pada pengukuran performa dua *Large Language Model* (LLM), yaitu LLaMA-2 dan GPT-3.5 Turbo, dalam tugas klasifikasi sentimen. Eksperimen dilakukan dengan membandingkan dua pendekatan, yaitu *Zero-shot* sebagai *baseline* dan *Retrieval-Augmented Few-Shot* (RAFS) sebagai metode utama yang diusulkan. Pada pendekatan *Zero-shot*, model melakukan klasifikasi sentimen tanpa diberikan contoh sebelumnya dalam prompt, sehingga performa yang diperoleh merepresentasikan kemampuan awal model dalam memahami instruksi klasifikasi [10]. Sebaliknya, pada pendekatan RAFS, model diberikan contoh-contoh relevan yang diperoleh melalui proses retrieval berbasis kemiripan semantik sebelum proses generasi dilakukan. Berikut adalah alur dari penelitian ini dapat dilihat pada Gambar 1.

Jurnal Sarjana Teknik Informatika
Vol. 14., No. 1, Februari 2026, pp. 42-52



Gambar 1 Alur Penelitian

Implementasi RAFS terdiri atas beberapa tahapan utama. Pertama, data ulasan diproses melalui tahap pra-pemrosesan yang meliputi *cleaning*, normalisasi, dan *case folding*. Selanjutnya, dataset dibagi menggunakan skema K-Fold Cross Validation untuk memastikan evaluasi yang adil dan menghindari bias. Pada setiap fold, data latih diubah menjadi representasi embedding dan disimpan dalam *vector database*. Ketika sebuah data uji diproses, sistem melakukan retrieval terhadap sejumlah k contoh terdekat berdasarkan *similarity score*. Contoh-contoh yang diperoleh kemudian digabungkan ke dalam prompt pada tahap augmentation, sebelum diberikan ke model LLM untuk proses generasi klasifikasi sentimen. Dengan skema ini, RAFS memungkinkan model menerima konteks tambahan yang relevan secara semantik, sehingga diharapkan dapat meningkatkan akurasi dibandingkan pendekatan *Zero-shot*. Pada Gambar 1, dijelaskan bahwa ada 2 alur yakni dengan *treatment* tambahan yakni RAFS dan *baseline* yakni *Zero-shot*

2.2 Akuisisi Data

Akuisisi data dilakukan melalui *Web Scrapping* untuk mengambil data mentah pada website. Pada penelitian ini, proses scraping dilakukan terhadap dua platform utama yaitu Google Review dan Booking.com, yang merupakan sumber ulasan pelanggan dengan relevansi tinggi terhadap objek studi kasus. Seluruh data yang diperoleh disimpan dalam format terstruktur CSV yang kemudian digunakan pada tahapan pelabelan dan Pra-pemrosesan lebih lanjut.

2.3 Pelabelan

Tahap pelabelan melibatkan 3 ahli yakni 2 Dosen Fakultas Bahasa dan Seni yaitu Kadek Trina Des Ryantini, S.Pd., M.Pd. dan Luh Putu Dian Kresnawati, S.Pd., M.Pd. selain itu terdapat 1 ahli dari pihak Hotel Blue Karma Ubud yaitu I Made Asta. Dari proses pelabelan didapat bahwa hanya ada beberapa perbedaan sentimen antara 3 ahli. Sebuah aturan diterapkan dalam menentukan sentimen final dari sebuah ulasan, jika jawaban kedua ahli dosen bahasa berbeda maka hasil final

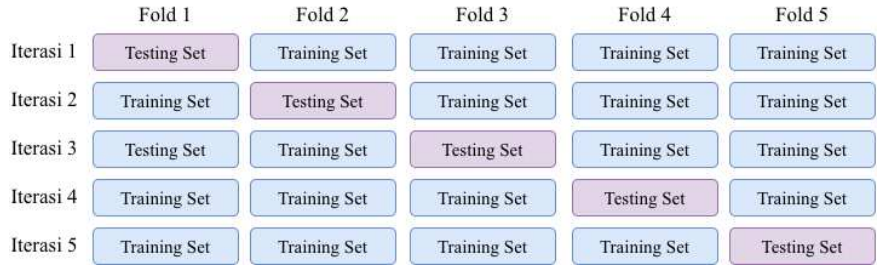
ditentukan oleh ahli dari pihak Blue Karma Ubud. Jika ketiga ahli menjawab sentimen yang berbeda pada satu ulasan, maka akan dilakukan diskusi lebih lanjut untuk penentuan akhir label. Dengan menggunakan mekanisme pelabelan berbasis konsensus ini, dataset yang digunakan dalam penelitian memiliki dasar anotasi yang lebih objektif dan dapat dipertanggungjawabkan secara ilmiah.

Guna memperkuat validitas proses tersebut secara statistik, tingkat konsistensi antar-ahli diukur menggunakan koefisien Fleiss' Kappa. Berdasarkan hasil pengujian terhadap total 669 ulasan, diperoleh nilai Kappa sebesar 0,5425 yang menunjukkan tingkat kesepakatan berada pada kategori *Moderate Agreement*. Nilai ini membuktikan bahwa meskipun terdapat variasi interpretasi bahasa, para ahli memiliki persepsi yang cukup konsisten dan objektif dalam menentukan kategori sentimen, sehingga label final yang dihasilkan layak digunakan sebagai ground truth untuk mengevaluasi model LLaMA-2 dan GPT-3.5 Turbo. Setelah dilakukan pelabelan, didapat bahwa sentimen positif memiliki persentase sebesar 81,8% dengan jumlah 547 ulasan, sedangkan sentimen netral sebesar 15,2% dengan jumlah 102 ulasan. Sementara itu, sentimen negatif memiliki persentase paling kecil, yaitu sekitar 3% dengan total 20 ulasan.

2.4 Pra-pemrosesan

Dari data *scraping* yang didapat masih terdapat beberapa *noise* maka dilakukan Pra-pemrosesan meliputi pembersihan teks dari simbol dan emoji, normalisasi kata tidak baku, serta penyeragaman huruf melalui *case folding*. Pra-pemrosesan yang dilakukan pada pendekatan RAFS lebih sederhana dibandingkan dengan pendekatan tradisional [4]. Proses ini bertujuan untuk mengurangi *noise* pada data teks sehingga representasi semantik ulasan menjadi lebih konsisten, selain itu fungsi dari Pra-pemrosesan ini agar data yang digunakan bersifat seragam.

2.5 Pembagian Data



Gambar 2 Viusalisasi K-fold Cross Validation

Pada penelitian ini digunakan pembagian data menggunakan *Stratified K-fold Cross Validation* yang bertujuan agar dataset tidak didistribusikan secara acak ke dalam *fold*, melainkan dengan tidak mengganggu rasio distribusi antar sampel kelas [11]. Berdasarkan hasil pembagian data, pada *fold* 1 dan *fold* 2 masing-masing terdapat 110 data sentimen positif, 20 data sentimen netral, dan 4 data sentimen negatif. Sementara itu, pada *fold* 3 dan 4 terdapat 109 data sentimen positif, 21 data sentimen netral, dan 4 data sentimen negatif, serta pada *fold* 5 terdapat 109 data sentimen positif, 20 data sentimen netral, dan 4 data sentimen negatif. Visual dari pembagian data dapat tertampil pada Gambar 2.

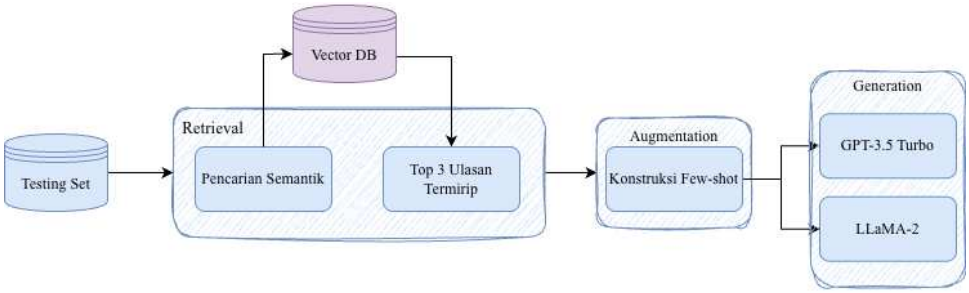
2.6 *Zero-shot*

Zero-shot dilakukan untuk mengukur kemampuan dasar dari kedua model dalam menganalisis sentimen tanpa diberikan contoh berlabel maupun mekanisme retrieval tambahan. Pada skenario ini, setiap ulasan yang menjadi data validasi langsung diberikan kepada model dalam bentuk prompt sederhana, sehingga proses klasifikasi sepenuhnya bergantung pada pengetahuan internal yang diperoleh model selama tahap pelatihan sebelumnya. Pendekatan *Zero-shot* digunakan untuk merepresentasikan kondisi awal model sebelum diterapkannya metode RAFS. Hasil dari skenario *Zero-shot* ini akan dijadikan *baseline* dan akan dikomparasikan dengan hasil setelah penerapan metode RAFS.

Jurnal Sarjana Teknik Informatika
Vol. 14., No. 1, Februari 2026, pp. 42-52

2.7 *Retrieval Augmented Few-shot*

Pada skenario *Zero-shot* model LLM akan diberikan data *testing* secara langsung dan hasil dari proses tersebut akan dilakukan evaluasi menggunakan *Confusion Matrix*. Sedangkan pada skema RAFS setiap iterasi saat data *testing* dijalankan maka akan dicari 3 ulasan yang memiliki kemiripan tinggi berdasarkan kedekatan semantik. Pada saat proses percarian ulasan, maka Vector DB secara otomatis mengubah data uji menjadi *embedding vector* dan akan dibandingkan dengan data *retrieval*. Setelah didapatkan 3 data termirip maka akan dilakukan proses augmentasi dengan menggabungkan contoh dan data testing untuk dijadikan *Few-shot* yang diberikan kepada model. Hasil dari jawaban model setelah mengklasifikasi menggunakan RAFS akan dievaluasi menggunakan *Confusion Matriks*. Dari Gambar 3, alur RAFS dijelaskan lebih rinci sebagai berikut:



Gambar 3 Alur Retrieval Augmented Few-shot

1) Input Embedding

Tahap pertama adalah melakukan embedding pada validasi set yang merupakan hasil dari pembagian pada tahap K-fold. Embedding dilakukan setiap ada data ulasan yang masuk ke dalam sistem RAFS. Proses embedding ini dilakukan setiap kali sebuah data ulasan validasi diproses oleh sistem RAFS. Ulasan tersebut akan direpresentasikan ke dalam bentuk vektor numerik agar dapat dibandingkan secara semantik dengan data pada retrieval set pada tahap selanjutnya

2) Retrieval

Query vector yang telah dihasilkan selanjutnya digunakan untuk melakukan pencarian kemiripan semantik pada Vector Database yang dibangun dari data retrieval. Proses pencarian ini menggunakan metrik Cosine Similarity untuk mengukur tingkat kedekatan antara query vector dengan seluruh vektor yang tersimpan di dalam database. Berdasarkan hasil perhitungan kemiripan tersebut, sistem kemudian mengambil tiga data dengan nilai kedekatan tertinggi terhadap data validasi dari knowledge base. Output dari tahap retrieval ini berupa ulasan-ulasan yang paling relevan secara kontekstual, yang selanjutnya dimanfaatkan sebagai *few-shot examples* pada tahap augmentasi

3) Agumentation

Pada tahap ini dilakukan proses penggabungan antara data ulasan hasil retrieval dan teks validasi berupa ulasan asli. Ulasan hasil retrieval berfungsi sebagai contoh *few-shot*, sedangkan ulasan validasi berperan sebagai input utama yang akan diklasifikasikan. Kedua komponen tersebut dirangkai ke dalam sebuah prompt final yang selanjutnya dikirimkan ke Large Language Model (LLM) sebagai konteks tambahan dalam proses inferensi. Ulasan yang telah terpilih pada proses retrieval kemudian digabungkan dengan satu data ulasan validasi untuk membentuk sebuah prompt final. Pada tahap ini, konsep few-shot memiliki peran penting karena menyediakan konteks dan pengetahuan awal yang relevan sehingga dapat mengarahkan proses penalaran model dalam menentukan sentimen ulasan validasi secara lebih akurat.

2.8 Evaluasi

Kinerja model GPT-3.5 Turbo dan LLaMA-2 dievaluasi dengan menghitung nilai akurasi sebagai ukuran performa keseluruhan, serta *confusion matrix* untuk menganalisis pola kesalahan klasifikasi pada tiap kelas sentimen. *Confusion matrix* adalah sebuah tabel yang digunakan untuk menghitung seberapa baik performa dari sistem klasifikasi. Berdasarkan penelitian oleh [12] mengenai analisis sentimen ditemukan bahwa penggunaan *confusion matrix* adalah seperti pada Tabel 1. Selain itu,

classification report digunakan untuk memperoleh nilai presisi, *recall*, dan *f1-score* pada setiap kelas, sehingga performa kedua model dapat dibandingkan secara objektif dan menyeluruh.

Tabel 1 *Confusion Matrix*

Confusion Matrix		Prediksi	
		Negatif	Positif
Aktual	Negatif	TN	FP
	Positif	FN	TP

Sehingga dari tabel *confusion matrix* tersebut dapat digunakan untuk menghitung akurasi, presisi, *recall* dan *f1-score*. Akurasinya adalah persentase jumlah prediksi yang benar dari total seluruh data uji dengan tujuan untuk mengukur kemampuan keseluruhan model dalam mengklasifikasikan ulasan dengan benar [13]. Berikut adalah rumus untuk menghitung akurasi ditunjukan pada Persamaan 1.

$$\text{Akurasi} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{ x } 100\% \quad \textbf{(1)}$$

Presisi mengukur tingkat ketepatan model dalam memprediksi suatu kelas tertentu, yaitu seberapa banyak prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dihasilkan model [14]. Nilai presisi yang tinggi menunjukkan bahwa model jarang menghasilkan prediksi positif yang salah atau *false positive*, sehingga penting pada kasus ketika kesalahan prediksi positif harus diminimalkan. presisi memiliki rumus ditunjukan pada Persamaan 2.

$$\text{Presisi} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \textbf{(2)}$$

Recall mengukur kemampuan model dalam menemukan seluruh data yang benar-benar termasuk dalam suatu kelas tertentu, yaitu perbandingan antara *true positive* dengan seluruh data yang seharusnya positif atau *true positive* dan *false negative* [15]. *Recall* yang tinggi menunjukkan bahwa model mampu menangkap sebagian besar data yang relevan, sehingga sangat penting ketika kehilangan data positif dianggap sebagai kesalahan serius. Rumus *recall* ditunjukan pada Persamaan 3.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \textbf{(3)}$$

F1-score merupakan ukuran keseimbangan antara presisi dan *recall* yang digunakan untuk menilai performa model secara lebih adil, terutama pada kondisi data tidak seimbang [16]. Nilai *f1-score* tinggi menandakan bahwa model tidak hanya akurat dalam memprediksi kelas tertentu, tetapi juga mampu menemukan sebagian besar data yang benar. Sementara itu, *f1-score weighted* mempertimbangkan jumlah data pada setiap kelas sehingga memberikan evaluasi performa model yang lebih representatif terhadap distribusi data secara keseluruhan. Berikut rumus dari *f1-score* ditunjukan pada Persamaan 4, sedangkan rumus dari *f1-score weighted* ditunjukan pada Persamaan 5.

$$\text{F1 Score} = 2 \text{ x } \frac{\text{Recall x Precision}}{\text{Recall} + \text{Precision}} \quad \textbf{(4)}$$

$$\text{F1}_{\text{weighted}} = \sum_{\text{i}=1}^{\text{n}} \left(\text{F1}_{\text{i}} \times \frac{\text{Support}_{\text{i}}}{\text{N}} \right) \quad \textbf{(5)}$$

2.9 Analisis Hasil

Untuk mengetahui seberapa signifikan RAFS mempengaruhi kedua model dilakukan analisis menggunakan Wilcoxon Signed-Rank. Uji ini dipilih karena metrik evaluasi yang dihasilkan dari proses *cross-validation* bersifat berpasangan dan tidak memenuhi asumsi normalitas. Uji Wilcoxon merupakan metode statistik non-parametrik yang sesuai untuk membandingkan dua metode pembelajaran mesin [17]. Tingkat signifikansi yang digunakan dalam penelitian ini adalah α=0.05.

Jurnal Sarjana Teknik Informatika
Vol. 14., No. 1, Februari 2026, pp. 42-52

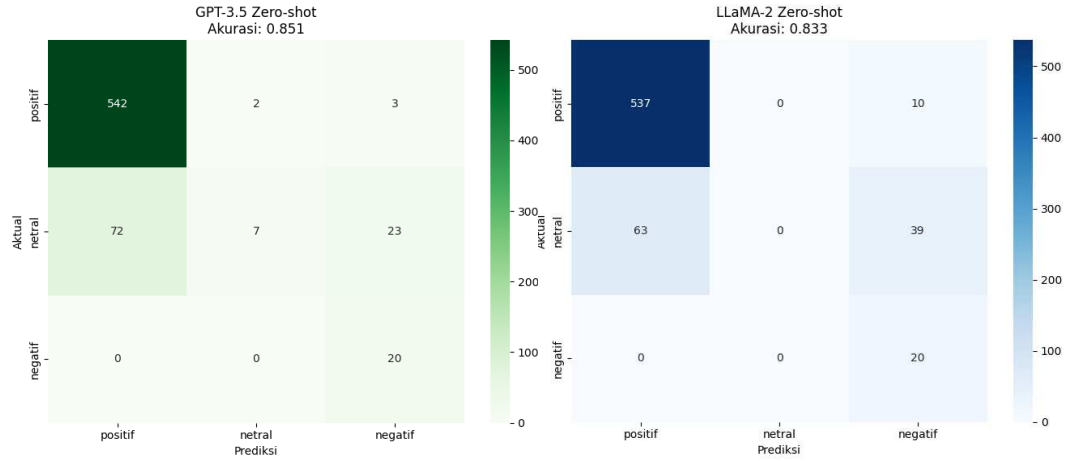
3. Hasil dan Pembahasan

3.1 Hasil

Evaluasi telah dilakukan terhadap 2 skenario yakni *Zero-shot* sebagai baseline pada kedua model LLM dan RAFS sebagai metode usulan untuk meningkatkan performa klasifikasi pada kedua model LLM.

3.1.1 Zero-shot

Pengujian *Zero-shot* diimplementasikan sebagai *baseline* untuk mengukur kemampuan awal kedua model tanpa bantuan konteks tambahan. Hasil evaluasi menunjukkan bahwa baik GPT-3.5 maupun LLaMA-2 menghadapi kendala signifikan dalam mengklasifikasikan sentimen netral, di mana LLaMA-2 tidak berhasil mengidentifikasi satu pun data pada kategori tersebut. Sebaliknya, kedua model menunjukkan performa yang sangat andal pada kelas positif serta kemampuan yang cukup konsisten dalam mengklasifikasikan sentimen negatif. Hasil dari *Confusion Matrix* pada skenario *Zero-shot* disajikan pada Gambar 4.



(a) *(b)*

Gambar 4 Hasil *Confusion Matrix* Zero-shot (a) GPT 3.5 Turbo (b) LLaMA-2

GPT-3.5 Turbo menunjukkan keunggulan pada seluruh metrik evaluasi dibandingkan LLaMA-2. Pada akurasi, GPT-3.5 Turbo mencapai skor 0,851, sementara LLaMA-2 berada di angka 0,833. Perbedaan performa yang paling signifikan terlihat pada metrik presisi, di mana GPT-3.5 Turbo mencatatkan skor 0,699 dibandingkan LLaMA-2 yang hanya sebesar 0,395. Untuk metrik *recall*, GPT-3.5 Turbo memperoleh 0,687 dan LLaMA-2 sebesar 0,661. Berdasarkan nilai presisi dan *recall* tersebut, GPT-3.5 Turbo menghasilkan *f1-score* sebesar 0,555, mengungguli LLaMA-2 yang memperoleh 0,462. Tren keunggulan ini juga tercermin pada nilai *weighted f1-score*, di mana GPT-3.5 Turbo memperoleh skor 0,801 dibandingkan LLaMA-2 sebesar 0,779. Hasil dari klasifikasi dari *Zero-shot* disajikan pada Tabel 2.

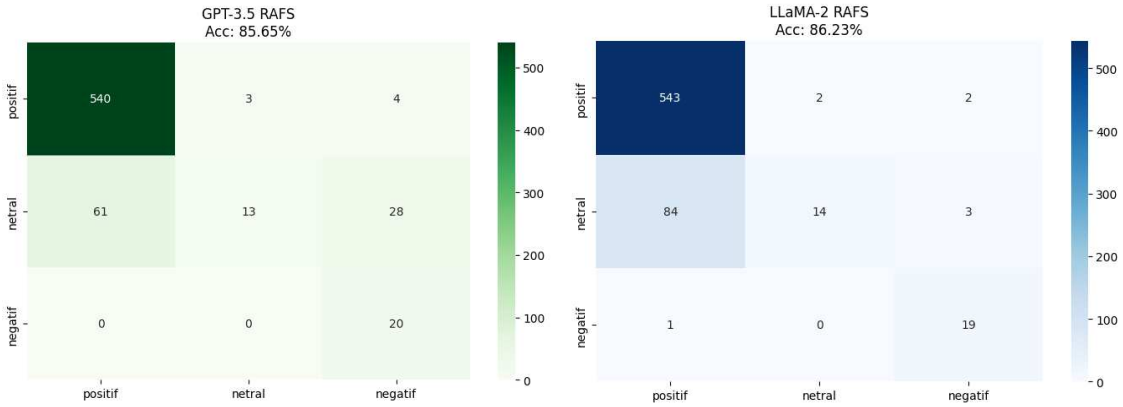
Tabel 2 Performa Zero-shot

Model	Akurasi	Presisi	Recall	F1-score	Weighted F1-Score
GPT 3.5 Turbo	0.851	0.699	0.687	0.555	0.801
LLaMA-2	0.833	0.395	0.661	0.462	0.779

3.1.2 *Retrieval Augmented Few-shot*

Pengujian metode RAFS sebagai metode usulan yang diimplementasikan untuk meningkatkan performa GPT-3.5 Turbo dan LLaMA-2. Hasil evaluasi menunjukkan bahwa kedua model mengalami peningkatan performa, khususnya pada klasifikasi sentimen netral. LLaMA-2, yang pada

pengujian *Zero-shot* gagal mengidentifikasi satu pun data pada kategori netral dengan 0 *true positive* (TP), menunjukkan peningkatan nyata dengan berhasil mengidentifikasi 14 data secara tepat melalui metode RAFS. Tren serupa terlihat pada GPT-3.5 yang peningkatannya tercatat dari 7 menjadi 13 data benar pada kelas yang sama. Sebaliknya, performa untuk kelas positif dan negatif menunjukkan stabilitas dengan hasil yang hampir identik antara skenario *baseline* dan metode usulan. Hasil *Confusion Matrix* disajikan pada Gambar 5.



(a) *(b)*

Gambar 5 Confusion Matriks Retrieval Augmented Few-shot (a) GPT 3.5 Turbo (b) LLaMA-2

Secara umum, LLaMA-2 menunjukkan keunggulan pada beberapa metrik evaluasi dibandingkan GPT 3.5 Turbo. Pada metrik akurasi, LLaMA-2 mencatatkan skor 0,862, melampaui GPT-3.5 Turbo yang memperoleh 0,856. Peningkatan paling mencolok terlihat pada presisi, di mana LLaMA-2 mencapai angka 0,844, sementara GPT-3.5 Turbo berada di angka 0,698. Meskipun GPT-3.5 Turbo memiliki nilai *recall* sedikit lebih tinggi yaitu 0,705 dibandingkan LLaMA-2 yang hanya memperoleh 0,694, LLaMA-2 tetap unggul secara agregat pada nilai *f1-score* dengan capaian 0,676 berbanding 0,572. Menariknya, kedua model menunjukkan performa yang identik pada metrik *weighted f1-score* dengan skor sebesar 0,819. Hasil dari klasifikasi dari *Zero-shot* disajikan pada Tabel 3.

Tabel 3 Performa Retrieval Augmented Few-shot

Model	Akurasi	Presisi	Recall	F1-score	Weighted F1-score
GPT 3.5 Turbo	0.856	0.698	0.705	0.572	0.819
LLaMA-2	0.862	0.844	0.694	0.676	0.819

3.1.3 Eksperimen Imbalance Class

Table 4 Akurasi Eksperimen *Imbalance Dataset*

	Oversampling	Non Oversampling
Akurasi GPT 3.5 Turbo	0.860	0.856
Akurasi LLaMA-2	0.855	0.862

Penelitian ini menghadapi tantangan ketidakseimbangan dataset yang signifikan, dengan dominasi sentimen positif sebesar 81,8%, diikuti oleh sentimen netral 15,2% dan negatif hanya 3%. Pada Tabel 4, eksperimen mitigasi menggunakan teknik oversampling pada kelas minoritas menunjukkan hasil yang tidak konsisten terhadap performa model, sementara akurasi GPT-3.5 Turbo meningkat marginal sebesar 1%, performa LLaMA-2 justru mengalami penurunan sebesar 0,7%.

e-ISSN: 2809-3399, p-ISSN: 2338-5197 Vol. 14., No. 1, Februari 2026, pp. 42-52

Temuan ini mengindikasikan bahwa bagi model LLM, penambahan frekuensi data secara artifisial melalui duplikasi tidak selalu efektif dalam meningkatkan kemampuan generalisasi, terutama pada arsitektur LLaMA-2 yang menunjukkan kecenderungan *overfitting* terhadap pola repetitif pada sampel minoritas. Hal ini menegaskan bahwa dalam konteks analisis sentimen ulasan pariwisata, keberagaman linguistik pada distribusi data asli jauh lebih krusial dibandingkan keseimbangan kuantitas antar kelas semata.

3.2 Pembahasan

Hasil eksperimen menunjukkan bahwa penerapan metode RAFS memberikan pengaruh yang signifikan, terutama pada model LLaMA-2. Temuan ini diperkuat oleh hasil uji statistik Wilcoxon pada Tabel 5, yang mencatatkan nilai *p-value* untuk LLaMA-2 sebesar 0.000074 yang dimana nilai tersebut < 0,05. Hal ini mengonfirmasi adanya perbedaan performa yang signifikan secara statistik setelah pemberian perlakuan. Sebaliknya, GPT-3.5 Turbo menunjukkan nilai *p-value* sebesar 0.2850 yang dimana > 0,05, yang mengindikasikan bahwa perubahan performa pada model tersebut tidak memiliki perbedaan yang signifikan secara statistik. Meskipun demikian, analisis lebih lanjut pada akurasi, presisi, *recall* dan *f1-score* menunjukkan bahwa kedua model mengalami perubahan ke arah positif, yang menandakan adanya peningkatan performa klasifikasi setelah implementasi metode RAFS.

Tabel 5 Hasil Uji Statistik Wilcoxon

Model	p-value	Kesimpulan
GPT-3.5 Turbo	0.2850	Tidak signifikan
LLaMA-2	0.000074	Signifikan

Dalam evaluasi pada skenario RAFS, model LLaMA-2 terbukti menunjukkan keunggulan yang lebih dominan dibandingkan GPT-3.5 Turbo pada mayoritas metrik utama. LLaMA-2 mencatatkan skor akurasi sebesar 0,862, melampaui GPT-3.5 Turbo yang berada pada angka 0,856. Perbedaan performa yang paling signifikan terlihat pada metrik presisi, di mana LLaMA-2 mencapai angka 0,844 sementara GPT-3.5 Turbo tertinggal jauh di angka 0,698. Keunggulan presisi yang tinggi ini berdampak langsung pada nilai *f1-score* LLaMA-2 yang mencapai 0,676, jauh mengungguli GPT-3.5 Turbo yang hanya memperoleh skor 0,572. Walaupun GPT-3.5 Turbo memiliki nilai *recall* yang sedikit lebih tinggi yakni 0,705 berbanding 0,694, secara agregat LLaMA-2 lebih unggul dalam memberikan hasil klasifikasi yang lebih tepat dan seimbang setelah diberikan augmentasi konteks.

Keunggulan LLaMA-2 pada skenario ini sekaligus menjawab kendala yang ditemukan pada tahap *Zero-shot*, di mana rendahnya nilai *f1-score* disebabkan oleh kecenderungan model memprediksi sentimen secara bias ke dalam kelas positif. Sebagaimana terlihat pada Gambar 3, banyak data netral yang salah diklasifikasikan sebagai positif karena keterbatasan model dalam memahami batasan sentimen tanpa panduan. Fenomena ini mengonfirmasi bahwa model open-source dengan skala parameter yang lebih kecil seperti LLaMA-2 sangat bergantung pada strategi tambahan seperti RAFS untuk mencapai performa optimal, terutama dalam menangani karakteristik dataset yang tidak seimbang. Ketangguhan LLaMA-2 dalam menangani ketidakseimbangan ini semakin dipertegas melalui uji tambahan menggunakan teknik *oversampling*. Berbeda dengan GPT-3.5 Turbo yang mengalami sedikit peningkatan akurasi dari 0,854 menjadi 0,864, LLaMA-2 justru menunjukkan performa terbaiknya pada distribusi data asli *non-oversampling* dengan akurasi sebesar 0,862 dan pada skenario oversampling menjadi 0.855. Penurunan performa LLaMA-2 menjadi 0,855 saat diterapkan oversampling mengindikasikan bahwa model ini lebih optimal dalam menggeneralisasi ulasan pariwisata melalui keberagaman linguistik pada data asli dibandingkan melalui replikasi data artifisial. Hal ini membuktikan bahwa mekanisme attention pada LLaMA-2, ketika didukung oleh metode RAFS, mampu mengidentifikasi fitur sentimen kelas minoritas secara efektif meskipun tanpa penyeimbangan jumlah sampel.

4. Kesimpulan

Penelitian ini berhasil mengimplementasikan metode RAFS pada model GPT-3.5 Turbo dan LLaMA-2 untuk tugas analisis sentimen pada domain pariwisata. Temuan eksperimen menunjukkan

bahwa penerapan RAFS memberikan kontribusi positif dalam meningkatkan berbagai metrik kedua model dibandingkan dengan skenario *baseline Zero-shot*. Dari eksperimen yang telah dilakukan didapat temuan bahwa RAFS pada kedua model berpengaruh dalam meningkatkan performa model, pada model LLaMA-2 RAFS berpengaruh secara signifikan. Sedangkan pada model GPT 3.5 Turbo RAFS tidak berpengaruh secara signifikan. Hal ini diperkuat dengan hasil uji statistik Wilcoxon yang menyatakan bahwa nilai *p-value* GPT-3.5 Turbo yaitu 0.2850 sedangkan LLaMA-2 0.000074, yang dimana jika nilai *p-value* kurang dari 0.05 maka dapat dikatakan ada perbedaan signifikan.

Perbandingan head-to-head antara kedua model setelah penerapan RAFS menunjukkan bahwa secara agregat LLaMA-2 sedikit lebih unggul dalam memberikan hasil klasifikasi yang presisi dan seimbang berkat dukungan augmentasi konteks. LLaMA-2 mengungguli GPT-3.5 Turbo pada metrik akurasi, presisi, dan *f1-score*, yang mencerminkan reliabilitas model dalam menangani distribusi data yang tidak seimbang. Sementara itu, GPT-3.5 Turbo hanya mencatatkan keunggulan tipis pada metrik recall, namun pencapaian tersebut diikuti dengan tingkat kesalahan prediksi (*false positive*) yang signifikan dibandingkan dengan LLaMA-2.

Deklarasi

Kontribusi Penulis. Semua penulis berkontrbusi secara bersama-sama dengan kontributor utama dalam artikel ini. Semua penulis membaca dan menyetujui versi akhir dari artikel yang diajukan.

Pernyataan Pendanaan. Tidak ada penulis yang menerima dana atau hibah dari lembaga atau badan pendanaan untuk penelitian ini.

Konflik Kepentingan. Penulis menyatakan tidak ada konflik kepentingan.

Informasi Tambahan. Tidak ada informasi tambahan dalam artikel ini.

Daftar Pustaka

[1] I. P. D. W. Darmawan, G. A. Pradnyana, and I. B. N. Pascima, "Optimasi Parameter Support Vector Machine Dengan Algoritma Genetika Untuk Analisis Sentimen Pada Media Sosial Instagram," SINTECH (Science and Information Technology) Journal, vol. 6, no. 1, pp. 58–67, Apr. 2023, doi: 10.31598/sintechjournal.v6i1.1245.

[2] W. Wijaya, K. A. Seputra, and N. P. N. P. Dewi, "Fine Tunning Model Indobert Untuk Analisis Sentimen Berita Pariwisata Indonesia," Jurnal Pendidikan Teknologi dan Kejuruan, vol. 22, no. 2, pp. 195–204, Jul. 2025, doi: 10.23887/jptk-undiksha.v22i2.104056.

[3] A. Y. Setiawan, I. Gede, M. Darmawiguna, and G. A. Pradnyana, "Sentiment Summarization Evaluasi Pembelajaran Menggunakan Algoritma Lstm (*Long Short Term Memory*)," Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika (KARMAPATI), vol. 11, no. 2, 2022.

[4] F. Nadi et al., "Sentiment Analysis Using Large Language Models: A Case Study of GPT-3.5," 2024, pp. 161–168. doi: 10.1007/978-981-97-0293-0_12.

[5] N. A. M. Herwanza, N. S. Harahap, F. Yanto, and F. Insani, "Penerapan Langchain Retriever dengan Model Chat Openai dalam Pengembangan Sistem Chatbot Hadis Berbasis Telegram," JTIM : Jurnal Teknologi Informasi dan Multimedia, vol. 6, no. 1, pp. 70–83, May 2024, doi: 10.35746/jtim.v6i1.514.

[6] P. Santra, M. Ghosh, D. Ganguly, P. Basuchowdhuri, and S. K. Naskar, "The 'Curious Case of Contexts' in Retrieval-Augmented Generation With a Combination of Labeled and Unlabeled Data," WIREs Data Mining and Knowledge Discovery, vol. 15, no. 2, Jun. 2025, doi: 10.1002/widm.70021.

[7] A. Rahman et al., "Comparative Analysis Based on DeepSeek, ChatGPT, and Google Gemini: Features, Techniques, Performance, Future Prospects," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2503.04783

[8] G. Yu, L. Liu, H. Jiang, S. Shi, and X. Ao, "Retrieval-Augmented Few-shot Text Classification," in Findings of the Association for Computational Linguistics: EMNLP 2023, Stroudsburg, PA, USA:

Association for Computational Linguistics, 2023, pp. 6721–6735. doi: 10.18653/v1/2023.findings-emnlp.447.

[9] L. Wang, N. Yang, and F. Wei, "Learning to Retrieve In-Context Examples for Large Language Models," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2307.07164

[10] Y. S. Yashwanth and R. Shettar, "Zero and Few Short Learning Using Large Language Models for De-Identification of Medical Records," IEEE Access, vol. 12, pp. 110385–110393, 2024, doi: 10.1109/ACCESS.2024.3439680.

[11] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross validation optimization on machine learning for prediction," Sinkron, vol. 7, no. 4, pp. 2407–2414, Oct. 2022, doi: 10.33395/sinkron.v7i4.11792.

[12] I. Kadek, Y. Prayoga, I. Made, G. Sunarya, and P. H. Suputra, "Klasifikasi Penyakit Demam Berdarah Menggunakan Algoritma Stacking Ensemble Learning", [Online]. Available: https://doi.org/10.31598

[13] P. W. Ariyani, I Made Gede Sunarya, and I Gede Aris Gunadi, "Analisis Sentimen Masyarakat Terhadap Virus Corona Berdasarkan Opini Dari Twitter Menggunakan Metode Naïve Bayes Dan K-Nearest Neighbor," Jurnal Pendidikan Teknologi dan Kejuruan, vol. 22, no. 2, pp. 128–138, Jul. 2025, doi: 10.23887/jptk-undiksha.v22i2.103233.

[14] I Putu Gede Hendra Suputra, Linawati, I. G. Sukadarmika, and N. P. Sastra, "Klasifikasi Judul Berita Bahasa Indonesia Menggunakan Support Vector Machine dan Seleksi Fitur Mutual Information," Jurnal Pendidikan Teknologi dan Kejuruan, vol. 22, no. 1, pp. 69–79, Jan. 2025, doi: 10.23887/jptkundiksha.v22i1.89158.

[15] N. K. T. A. Saputri, I. G. A. Gunadi, and I. M. G. Sunarya, "Analisis Sentimen Pelayanan Daring di Fakultas Teknik dan Kejuruan Universitas Pendidikan Ganesha Menggunakan Algoritma Naïve Bayes dan LSTM," MALCOM: Indonesian Journal of Machine Learning and Computer Science, vol. 4, no. 3, pp. 1120–1129, Jul. 2024, doi: 10.57152/malcom.v4i3.1336.

[16] P. S. Ghatora, S. E. Hosseini, S. Pervez, M. J. Iqbal, and N. Shaukat, "Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM," Big Data and Cognitive Computing, vol. 8, no. 12, p. 199, Dec. 2024, doi: 10.3390/bdcc8120199.

[17] K. Džermeikaitė, J. Krištolaitytė, and R. Antanaitis, "Application of Machine Learning Models for the Early Detection of Metritis in Dairy Cows Based on Physiological, Behavioural and Milk Quality Indicators," Animals, vol. 15, no. 11, p. 1674, Jun. 2025, doi: 10.3390/ani15111674.