

BAB I

PENDAHULUAN

1.1 Latar Belakang

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. Hal ini karena ulasan pelanggan mencerminkan persepsi, pengalaman, serta tingkat kepuasan konsumen terhadap suatu produk atau layanan. Salah satu industri akomodasi yang memperhatikan ulasan pelanggan adalah Blue Karma Ubud, hotel ini merupakan akomodasi yang terletak di Ubud dengan mengedepankan citra premium berbasis kenyamanan dan keindahan alam. Ulasan menjadi salah satu bahan evaluasi yang digunakan oleh Blue Karma Ubud untuk meningkatkan kualitas pelayanannya. Idealnya pihak Blue Karma Ubud dapat merespon dengan cepat ulasan yang tersebar diberbagai platform. Namun dalam menganalisis ulasan pihak Blue Karma Ubud masih menggunakan metode tradisional dengan membaca satu persatu ulasan tanpa menggunakan sebuah pendekatan yang efisien. Kesenjangan antara kebutuhan akan wawasan yang cepat dan metode analisis yang lambat yang menjadi permasalahan utama pihak Blue Karma Ubud. Untuk menjembatani permasalahan ini diperlukan solusi otomatis yakni analisis sentimen.

Menurut (Darmawan et al., 2023) analisis sentimen adalah bidang studi yang berfokus menganalisis opini, sentimen, emosi, penilaian, dan sikap seseorang terhadap suatu topik. Analisis sentimen telah banyak digunakan oleh perusahaan

untuk mengetahui review dari produk atau jasa yang ditawarkan (Setiawan et al., 2022). Analisis sentimen dapat dilakukan dengan memanfaatkan *machine learning* tradisional seperti *Support Vector Machine (SVM)*, *Naive Bayes*, atau *Random Forest*. Namun, menurut penelitian Nadi et al. (2024) menggunakan metode tradisional untuk pengolahan data teks menghadapi beberapa tantangan, terutama dalam hal akurasi dan pemahaman berbagai bahasa. 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 (Herwanza et al., 2024). LLM dibangun dengan arsitektur Transformer yang memungkinkan model memahami hubungan semantik antar kata. Salah satu keunggulan LLM dalam analisis teks adalah kemampuannya untuk menyesuaikan diri dengan berbagai domain bahasa, termasuk bahasa informal yang umum digunakan serta hubungan antar kalimat yang kompleks, sehingga mampu menghasilkan interpretasi yang lebih akurat dan mendalam. Dalam tugas analisis sentimen LLM dapat dikombinasikan dengan berbagai metode, seperti *Zero-shot*, *Fine Tuning*, *Few-shot* atau *Retrieval Augmented Generation (RAG)*.

Dari berbagai pendekatan yang dapat diterapkan pada LLM, pemilihan metode yang tepat menjadi sangat krusial. Meskipun LLM dapat memahami teks dengan baik, pada kenyataannya LLM dapat mengalami halusinasi terutama dalam hal mengakses informasi terkini atau spesifik yang tidak terdapat dalam data pelatihannya (Tonmoy et al., 2024). 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 Santra et al. (2025) membuktikan bahwa performa RAG dapat ditingkatkan lebih lanjut lagi, terutama dengan memberikan penalaran berbasis contoh. Selain itu penelitian oleh Rahman et al., (2025) menunjukkan bahwa hasil dari metode Fewshot sangat bergantung pada relevansi dari contoh yang dipilih. Kedua penelitian tersebut menunjukkan 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 (Yu et al., 2023). Pendekatan ini mengadopsi kemampuan RAG dalam pengambilan informasi yang relevan yang disebut *retrieval* dan pendekatan *Few-shot* digunakan untuk mengarahkan penalaran model dilakukan pemberian contoh berupa ulasan dan labelnya secara eksplisit langsung didalam prompt. Sistem serupa telah dikembangkan oleh Wang et al. (2024) 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 karena kedua model ini memiliki dokumentasi arsitektur LLM yang lengkap. 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 *Retrival Augmented Few-shot*. Pernyataan ini didukung oleh penelitian Buscemi & Proverbio (2024) yang menunjukkan bahwa hasil klasifikasi dua model yang berbeda memiliki hasil berbeda. Selain itu penelitian oleh Gorriz et al. (2024) menegaskan bahwa bahkan dengan dataset yang sama, performa dari satu penelitian dapat berbeda secara substansial dengan penelitian lain. Kedua faktor ini menyoroti adanya celah penelitian, berupa minimnya penelitian yang menyajikan perbandingan *head-to-head* yang secara empiris mengukur performa kedua LLM LLaMA 2 dan GPT-3.5 Turbo dalam sebuah arsitektur *Retrival Augmented Few-shot* yang terkontrol. Oleh sebab itu peneliti mengusulkan sebuah penelitian berjudul **“Perbandingan Model GPT-3.5 Turbo Dan LLaMA-2 Dalam Analisis Sentimen Ulasan Pada Blue Karma Ubud”**. 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.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dapat diidentifikasi beberapa permasalahan sebagai berikut:

1. Model LLM beresiko mengalami halusinasi dan kesulitan memahami ulasan yang kompleks.
2. Dalam konteks analisis sentimen RAG tidak hanya cukup diberi konteks tanpa label.
3. Hasil dari metode *Few-shot* sangat bergantung pada relevansi dari contoh yang dipilih.
4. GPT dan LLaMA memiliki performa yang tidak sama walau dihadapkan tugas yang sama.

1.3 Rumusan Masalah

1. Bagaimana kombinasi metode *Retrieval-Augmented Generation* (RAG) dengan *Few-shot* dapat digunakan dalam tugas analisis sentimen?
2. Bagaimana penerapan arsitektur *Retrieval-Augmented Few-Shot* dalam mendukung proses analisis sentimen?
3. Bagaimana perbandingan performa antara model GPT-3.5 Turbo dan LLaMA-2 ketika diimplementasikan dalam arsitektur *Retrieval-Augmented Few-shot* untuk tugas analisis sentimen?

1.4 Tujuan

1. Untuk mengkombinasikan metode *Retrieval-Augmented Generation* (RAG) dengan *Few-Shot* dapat digunakan dalam tugas analisis sentimen.

2. Untuk menerapkan arsitektur *Retrieval-Augmented Few-shot* dalam mendukung proses analisis sentimen.
3. Menentukan dan merekomendasikan model terbaik antara GPT-3.5 Turbo dan LLaMA 2 yang menunjukkan performa paling unggul dan andal untuk diimplementasikan dalam sistem analisis sentimen menggunakan metode *Retrieval-Augmented Few-shot*.

1.5 Batasan Masalah

Agar penelitian ini terarah dan fokus, maka ditetapkan pembatasan masalah sebagai berikut:

1. Data yang digunakan hanya data ulasan Hotel Blue Karma Ubud pada platform Booking.com dan Google Review.
2. Analisis hanya difokuskan pada isi ulasan berbasis teks, tanpa mempertimbangkan ulasan berbentuk gambar atau video.
3. Sentimen yang digunakan hanya sentimen positif, netral dan negatif dari rentang tahun 2015 bulan Januari sampai 2024 bulan Desember.
4. Hanya fokus pada penerapan RAFS pada model GPT-3.5 Turbo dan LLaMA 2.
5. Website yang dibuat hanya untuk membuktikan bahwa skenario terbaik dapat diterapkan, tidak ada analisis dan pengembangan lebih lanjut.

1.6 Manfaat

1. Manfaat Teoritis
 - a. Memberikan kontribusi dalam bidang *Natural Language Processing* (NLP), khususnya dalam tugas analisis sentimen dengan menggunakan

model *Large Language Model* (LLM) seperti LLaMA-2 dan GPT 3.5 Turbo.

- b. Menambah referensi ilmiah terkait efektivitas *Retrival Augmented Few-shot* dalam analisis sentimen, sehingga dapat menjadi acuan bagi penelitian di masa mendatang.

2. Manfaat Praktis

- a. Menyediakan kajian perbandingan antara LLaMA-2 dan GPT 3.5 Turbo dalam analisis sentimen ulasan pelanggan, yang dapat membantu peneliti memahami keunggulan dan keterbatasan masing-masing model dalam memahami pola sentimen pelanggan lebih efektif
- b. Memberikan rekomendasi mengenai model dan metode yang paling tepat dalam menganalisis ulasan pelanggan, yang dapat diterapkan oleh bisnis lain dalam industri perhotelan

