Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 5181 – 5191 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

PERBANDINGAN KESESUAIAN PENDAPAT MANUSIA DAN AI (CHATGPT, GEMINI, DEEPSEEK) MENGGUNAKAN PENDEKATAN GROUND TRUTH

Kelvin[1], Sunaryo Winardi[2], Handoko[3], Erwin Setiawan Panjaitan[4], Rivaldi Lubis[5]
Universitas Mikroskil, Medan
Email: [1]kelvin.chen@mikroskil.ac.id, [2]sunaryo.winardi@mikroskil.ac.id,
[3]handoko.wu@mikroskil.ac.id, [4]erwin@mikroskil.ac.id, [5]rivaldi.lubis@mikroskil.ac.id

Abstract: *The rapid advancement of Large Language Models (LLMs) has significantly improved sentiment analysis capabilities. However, the extent to which these models produce sentiment classifications consistent with human judgment remains an important research question. This study aims to evaluate and compare the agreement of ChatGPT, Gemini, and DeepSeek with human-generated ground truth in sentiment analysis. A total of 1,497 product reviews were collected from the Sephora Products and Skincare Reviews dataset. Three independent annotators labeled each review as positive, neutral, or negative to establish the ground truth. The annotation reliability achieved a Fleiss' Kappa coefficient of 0.7053, indicating substantial agreement and confirming the reliability of the ground truth for evaluation purposes. Subsequently, the three LLMs performed sentiment classification using an Expectation–Role–Action (ERA) prompting strategy. Model performance was assessed using Cohen's Kappa to measure agreement with the ground truth and Spearman's rank correlation to evaluate the consistency of sentiment rankings. The results show that DeepSeek achieved the highest performance, with an average Cohen's Kappa of 0.753 and an average Spearman's correlation coefficient of 0.860, followed by Gemini ($\kappa = 0.662$; $\rho = 0.834$), while ChatGPT demonstrated the lowest agreement ($\kappa = 0.223$; $\rho = 0.367$). These findings indicate that the three LLMs exhibit significantly different levels of agreement with human judgment, with DeepSeek producing sentiment classifications that most closely align with the established ground truth.*

Keywords: *Large Language Models, Sentiment Analysis, Ground Truth, Cohen's Kappa, Spearman's Rank Correlation.*

Abstrak: Perkembangan Large Language Models (LLMs) telah meningkatkan kemampuan analisis sentimen berbasis kecerdasan buatan. Namun, tingkat kesesuaian hasil klasifikasi sentimen yang dihasilkan oleh berbagai LLM terhadap penilaian manusia masih memerlukan evaluasi yang komprehensif. Penelitian ini bertujuan membandingkan tingkat kesesuaian hasil analisis sentimen ChatGPT, Gemini, dan DeepSeek terhadap ground truth yang diperoleh melalui anotasi manusia. Penelitian menggunakan 1.497 ulasan produk dari dataset Sephora Products and Skincare Reviews. Sebanyak tiga anotator independen melakukan pelabelan sentimen ke dalam kategori positif, netral, dan negatif untuk membentuk ground truth. Hasil pengujian reliabilitas menunjukkan nilai Fleiss' Kappa sebesar 0,7053, yang mengindikasikan tingkat kesepakatan Substantial Agreement, sehingga ground truth layak digunakan sebagai acuan evaluasi. Selanjutnya, ketiga model LLM melakukan klasifikasi sentimen menggunakan prompt berbasis Expectation–Role–Action (ERA). Tingkat kesesuaian hasil klasifikasi dievaluasi menggunakan Cohen's Kappa, sedangkan konsistensi hubungan dengan ground truth dianalisis menggunakan korelasi Spearman. Hasil penelitian menunjukkan bahwa DeepSeek memberikan performa terbaik dengan rata-rata Cohen's Kappa sebesar 0,753 dan rata-rata korelasi Spearman sebesar 0,860, diikuti oleh Gemini ($\kappa = 0{,}662$; $\rho = 0{,}834$), sedangkan ChatGPT memperoleh tingkat kesesuaian terendah ($\kappa = 0{,}223$; $\rho = 0{,}367$). Temuan ini menunjukkan bahwa terdapat perbedaan kemampuan interpretasi sentimen antar model LLM, dengan DeepSeek menghasilkan klasifikasi yang paling mendekati penilaian manusia pada dataset yang digunakan.

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 5181 – 5191 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Kata kunci: Large Language Model, Analisis Sentimen, Ground Truth, Cohen's Kappa, Spearman Correlation.

PENDAHULUAN

Perkembangan Artificial Intelligence (AI), khususnya Large Language Models (LLMs), telah membawa kemajuan yang signifikan dalam bidang Natural Language Processing (NLP). Model seperti ChatGPT, Gemini, dan DeepSeek memiliki kemampuan memahami konteks bahasa alami, menghasilkan teks yang menyerupai tulisan manusia, serta menyelesaikan berbagai tugas berbasis bahasa, seperti penerjemahan, question answering, peringkasan dokumen, dan analisis sentimen [1]–[4]. Kemampuan tersebut mendorong pemanfaatan LLM pada berbagai sektor, termasuk pendidikan, bisnis, kesehatan, dan layanan publik, karena mampu meningkatkan efisiensi pengolahan informasi berbasis teks.

Salah satu penerapan penting LLM adalah analisis sentimen, yaitu proses mengidentifikasi dan mengklasifikasikan opini pengguna ke dalam kategori positif, netral, atau negatif. Dibandingkan metode klasifikasi konvensional, LLM memiliki kemampuan memahami hubungan semantik dan konteks kalimat yang lebih baik sehingga berpotensi menghasilkan interpretasi sentimen yang lebih akurat [5]-[7]. Namun demikian, kemampuan generatif yang tinggi tidak selalu menjamin bahwa hasil klasifikasi sentimen yang dihasilkan telah sesuai dengan persepsi manusia. Berbagai penelitian menunjukkan bahwa LLM masih dapat menghasilkan hallucination, bias interpretasi, maupun inkonsistensi dalam memahami konteks bahasa tertentu. Kondisi tersebut menjadi tantangan, terutama pada aplikasi yang membutuhkan hasil analisis sentimen yang mampu merepresentasikan penilaian manusia secara konsisten [8]-[12].

Dalam evaluasi model klasifikasi, ground truth yang diperoleh melalui anotasi manusia merupakan acuan utama untuk menilai kualitas prediksi model [11]-[13]. Berbeda dengan evaluasi yang hanya menggunakan metrik klasifikasi konvensional, pendekatan berbasis human ground truth memungkinkan penilaian yang lebih representatif terhadap kemampuan model dalam memahami makna dan konteks suatu opini. Oleh karena itu, tingkat kesesuaian antara hasil klasifikasi model dan ground truth manusia menjadi indikator penting untuk menilai keandalan LLM dalam tugas analisis sentiment [13].

Beberapa penelitian sebelumnya telah mengevaluasi performa LLM pada berbagai domain, seperti pendidikan, kesehatan, dan biomedical natural language processing [14]–[15]. Namun, sebagian besar penelitian tersebut berfokus pada pengukuran akurasi atau perbandingan performa dengan model pembelajaran mesin lainnya. Penelitian yang secara khusus membandingkan tingkat kesesuaian ChatGPT, Gemini, dan DeepSeek terhadap ground truth yang dibangun melalui anotasi manusia pada tugas analisis sentimen masih relatif terbatas. Selain itu, evaluasi umumnya hanya menggunakan satu metrik sehingga belum mampu menggambarkan tingkat kesepakatan maupun konsistensi hubungan antara hasil klasifikasi model dan penilaian manusia secara komprehensif. Kondisi tersebut menunjukkan adanya research gap dalam evaluasi kemampuan LLM untuk merepresentasikan persepsi manusia pada analisis sentimen [11], [13], [16]–[18].

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan membandingkan tingkat kesesuaian hasil analisis sentimen yang dihasilkan oleh ChatGPT, Gemini, dan DeepSeek terhadap ground truth yang dibentuk melalui anotasi tiga evaluator independen menggunakan 1.497 ulasan dari dataset Sephora Products and Skincare Reviews [19]. Reliabilitas ground truth terlebih dahulu dievaluasi menggunakan Fleiss' Kappa, sedangkan

tingkat kesesuaian masing-masing model terhadap ground truth dianalisis menggunakan Cohen's Kappa dan Spearman's Rank Correlation [13],[20]. Penelitian ini diharapkan memberikan kontribusi berupa evaluasi komparatif tiga LLM populer menggunakan human ground truth yang tervalidasi, sehingga dapat menjadi referensi dalam pemilihan model LLM untuk aplikasi analisis sentimen yang memerlukan interpretasi mendekati penilaian manusia [11], [13].

TINJAUAN PUSTAKA

Large Language Models dalam Analisis Sentimen

Large Language Models (LLMs) merupakan model berbasis arsitektur Transformer yang dilatih menggunakan korpus teks berskala besar sehingga mampu memahami hubungan semantik, konteks, dan pola bahasa secara lebih komprehensif dibandingkan pendekatan machine learning konvensional [1]–[4]. Kemampuan tersebut memungkinkan LLM menyelesaikan berbagai tugas Natural Language Processing (NLP), seperti question answering, penerjemahan, peringkasan dokumen, hingga analisis sentimen [2]–[7].

Dalam analisis sentimen, LLM mampu menginterpretasikan opini berdasarkan konteks kalimat tanpa memerlukan proses *feature engineering* yang kompleks [5]–[7]. Beberapa model yang banyak digunakan saat ini antara lain ChatGPT, Gemini, dan DeepSeek [3], [16]–[18]. Ketiga model tersebut memiliki karakteristik yang berbeda dalam memahami konteks bahasa, sehingga berpotensi menghasilkan interpretasi sentimen yang berbeda terhadap data yang sama [16]–[18].

Ground Truth dan Evaluasi Kesesuaian

Dalam *machine learning*, **ground truth** merupakan label acuan yang dianggap benar untuk mengevaluasi kinerja model dan umumnya diperoleh melalui proses anotasi manual oleh manusia. Pada analisis sentimen berbasis *Large Language Model* (LLM), penggunaan *ground truth* menjadi penting karena meskipun model mampu menghasilkan keluaran yang koheren secara linguistik, hasilnya belum tentu selaras dengan persepsi manusia [8][21]. Oleh karena itu, praktik terbaik dalam penyusunan *ground truth* menerapkan skema multi-anotator untuk mengurangi subjektivitas individu, sedangkan reliabilitas hasil anotasi dievaluasi menggunakan metrik kesepakatan antarpenilai, seperti Cohen's Kappa atau Fleiss' Kappa, guna memastikan bahwa dataset yang digunakan sebagai acuan memiliki tingkat konsistensi yang memadai [22].

Kesesuaian antara hasil klasifikasi sentimen model AI dan *ground truth* manusia dapat dievaluasi menggunakan beberapa metrik statistik. **Cohen's Kappa (κ)** digunakan untuk mengukur tingkat kesepakatan antara dua penilai di luar kemungkinan kesepakatan yang terjadi secara acak, dengan nilai yang semakin mendekati 1 menunjukkan tingkat kesepakatan yang semakin tinggi [22]. **Uji Chi-square (χ^2)** digunakan untuk menguji apakah terdapat perbedaan distribusi hasil klasifikasi antara model AI dan *ground truth*, di mana nilai χ^2 yang rendah menunjukkan distribusi yang semakin serupa [23]. Sementara itu, **korelasi Spearman (ρ)** digunakan untuk mengukur keselarasan hubungan monotonik antara hasil klasifikasi AI dan penilaian manusia pada data ordinal, dengan nilai ρ yang semakin mendekati 1 menunjukkan kesesuaian interpretasi sentimen yang semakin kuat [20]. Persamaan yang digunakan untuk menghitung masing-masing metrik ditunjukkan pada Persamaan (1), Persamaan (2), dan Persamaan (3).

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

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

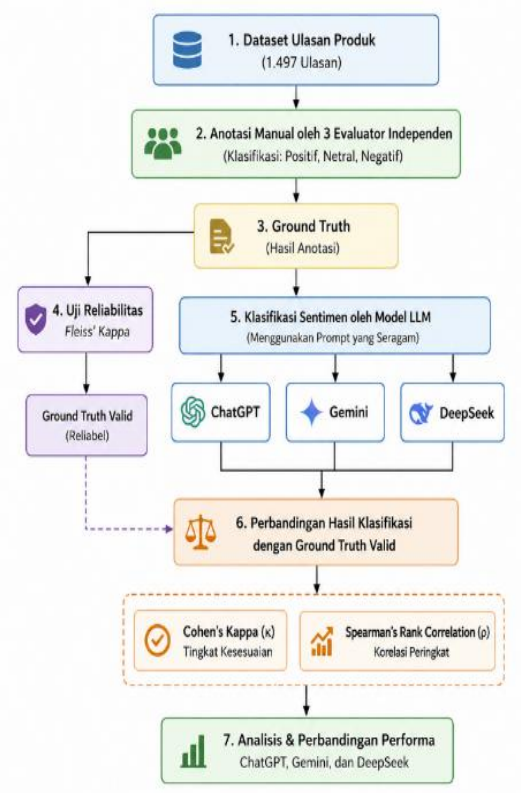
$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (3)$$

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 5181 – 5191 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

METODE

Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif komparatif untuk mengevaluasi tingkat kesesuaian hasil analisis sentimen yang dihasilkan oleh tiga Large Language Models (LLMs), yaitu ChatGPT, Gemini, dan DeepSeek, terhadap ground truth yang dibentuk melalui anotasi manusia. Proses penelitian diawali dengan pembentukan ground truth melalui anotasi tiga evaluator independen, kemudian setiap model melakukan klasifikasi sentimen terhadap dataset yang sama menggunakan prompt yang seragam. Selanjutnya, reliabilitas ground truth dievaluasi menggunakan Fleiss' Kappa, sedangkan tingkat kesesuaian hasil klasifikasi masing-masing model terhadap ground truth dianalisis menggunakan Cohen's Kappa dan Spearman's Rank Correlation. Ada pun desain penelitian dapat dilihat pada Gambar 1, berikut ini.



Gambar 1 Alur Penelitian Evaluasi Kesesuaian Hasil Analisis Sentimen Model LLM terhadap *Ground Truth Manusia*

Dataset

Dataset yang digunakan berasal dari Sephora Products and Skincare Reviews yang tersedia pada platform Kaggle [19]. Dataset ini berisi ulasan pelanggan terhadap berbagai produk perawatan kulit dan kosmetik. Sebanyak 1.497 ulasan dipilih secara acak sebagai sampel penelitian. Seluruh ulasan digunakan dalam bentuk asli tanpa mengubah isi teks agar proses anotasi dan evaluasi mencerminkan kondisi data yang sebenarnya.

Pembentukan Ground Truth

Ground truth dibentuk melalui proses anotasi manual oleh tiga evaluator independen. Setiap evaluator memberikan label sentimen terhadap seluruh ulasan ke dalam tiga kategori, yaitu positif, netral, dan negatif. Untuk memastikan konsistensi hasil anotasi, dilakukan pengujian reliabilitas menggunakan Fleiss' Kappa. Nilai koefisien yang diperoleh digunakan untuk menentukan kelayakan ground truth sebagai acuan dalam proses evaluasi model.

Implementasi Large Language Models

Penelitian ini mengevaluasi tiga model LLM, yaitu ChatGPT, Gemini, dan DeepSeek. Ketiga model diberikan dataset yang sama serta menggunakan prompt berbasis Expectation–Role–Action (ERA) dengan instruksi yang identik. Setiap model diminta mengklasifikasikan setiap ulasan ke dalam salah satu dari tiga kategori sentimen, yaitu positif, netral, atau negatif, tanpa memberikan penjelasan tambahan. Penggunaan prompt yang seragam bertujuan meminimalkan variasi keluaran yang disebabkan oleh perbedaan instruksi, sehingga perbandingan antar model dilakukan pada kondisi yang setara.

Metode Evaluasi

Evaluasi dilakukan menggunakan tiga ukuran statistik yang saling melengkapi. Fleiss' Kappa digunakan untuk mengukur reliabilitas hasil anotasi

tiga evaluator dalam membentuk ground truth. Cohen's Kappa digunakan untuk mengukur tingkat kesesuaian antara hasil klasifikasi masing-masing model dengan ground truth setelah memperhitungkan kemungkinan kesepakatan yang terjadi secara acak. Selanjutnya, Spearman's Rank Correlation digunakan untuk mengukur konsistensi hubungan antara hasil klasifikasi model dan ground truth.

Sebagai analisis pendukung, penelitian ini juga menerapkan uji Chi-Square untuk mengidentifikasi apakah distribusi hasil klasifikasi yang dihasilkan oleh masing-masing model berbeda secara signifikan dibandingkan dengan distribusi ground truth manusia. Kombinasi keempat metode tersebut memberikan evaluasi yang komprehensif terhadap reliabilitas ground truth, tingkat kesepakatan hasil klasifikasi, konsistensi hubungan, serta karakteristik distribusi prediksi setiap model.

HASIL DAN PEMBAHASAN

Pembentukan *Ground Truth*

Pelabelan data dilakukan secara manual oleh tiga anotator independen yang memiliki pemahaman mengenai analisis sentimen dan konteks ulasan produk. Sebanyak 1.497 ulasan diklasifikasikan ke dalam tiga kategori sentimen, yaitu positif, netral, dan negatif, untuk membentuk *ground truth* yang digunakan sebagai acuan dalam pelatihan dan evaluasi model analisis sentimen. Distribusi hasil pelabelan oleh masing-masing anotator disajikan pada **Tabel 1**.

Tabel 1. Distribusi Label Sentimen oleh Tiga Anotator

Kategori Sentimen	Anotator 1	Anotator 2	Anotator 3
Positif	**1.183**	**1.124**	**1.153**
Netral	**171**	**215**	**175**
Negatif	**143**	**158**	**169**
Total	**1.497**	**1.497**	**1.497**

Untuk memastikan reliabilitas hasil pelabelan dan mengukur tingkat konsistensi antarpenilai, dilakukan pengujian menggunakan **Fleiss' Kappa**, yaitu metode statistik untuk mengukur *inter-rater agreement* pada lebih dari dua anotator dengan kategori nominal. Pada penelitian ini, jumlah ulasan yang dinilai adalah **1.497** (N = 1.497), dengan **tiga anotator** (n = 3) dan **tiga kategori sentimen** (k = 3), yaitu positif, netral, dan negatif.

Perhitungan Fleiss' Kappa diawali dengan menentukan tingkat kesepakatan setiap ulasan (P_i), yaitu **1** jika ketiga anotator memberikan label yang sama, **1/3** jika dua anotator sepakat, dan **0** jika seluruh anotator memberikan label yang berbeda. Rata-rata seluruh nilai P_i menghasilkan nilai kesepakatan aktual ($\bar{P}$) sebesar **0,887998219**, yang menunjukkan tingkat konsistensi pelabelan yang tinggi antar anotator. Selanjutnya, proporsi setiap kategori sentimen (P_j) dihitung dari keseluruhan penilaian, yaitu **0,770429748** untuk kategori positif, **0,118681808** untuk kategori netral, dan **0,110888444** untuk kategori negatif. Berdasarkan proporsi tersebut diperoleh nilai kesepakatan yang diharapkan (P_e) sebesar **0,619943616**. Nilai Fleiss' Kappa kemudian dihitung:

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e} = \frac{0{,}8879 - 0{,}6199}{1 - 0{,}6199} = 0{,}7053$$

Nilai Fleiss' Kappa yang diperoleh adalah **κ = 0,7053**, yang termasuk dalam kategori **Substantial Agreement (0,61–0,80)** berdasarkan interpretasi Landis dan Koch[24]. Nilai ini mencerminkan **tingkat kesepakatan yang tinggi antar-anotator**, menunjukkan bahwa para penilai memiliki persepsi yang relatif seragam terhadap ekspresi sentimen dalam data yang dianalisis.

Evaluasi Hasil Klasifikasi LLM

Distribusi hasil klasifikasi sentimen yang dihasilkan oleh ChatGPT, Gemini, dan DeepSeek ditunjukkan pada Tabel 2.

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 5181 – 5191 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Tabel 2. Distribusi hasil prediksi dari ketiga model

Model LLM	Positif	Netral	Negatif	Total
ChatGPT	1171	260	66	1497
Gemini	1179	64	254	1497
DeepSeek	1151	159	187	1497

Secara umum, ketiga model berhasil mengidentifikasi dominasi sentimen positif pada dataset. Namun, terdapat perbedaan karakteristik dalam mengklasifikasikan sentimen netral dan negatif. ChatGPT menghasilkan proporsi prediksi positif yang lebih tinggi dibandingkan model lainnya, sedangkan Gemini menghasilkan jumlah prediksi negatif yang relatif lebih banyak. DeepSeek menghasilkan distribusi prediksi yang paling mendekati distribusi ground truth, menunjukkan kemampuan yang lebih seimbang dalam membedakan ketiga kategori sentimen.

Tingkat kesesuaian hasil klasifikasi terhadap ground truth kemudian dievaluasi menggunakan Cohen's Kappa. Hasil pengujian disajikan pada **Tabel 3.**

Tabel 3. Nilai Cohen's Kappa (κ) antara Model AI dan Anotator Manusia

Model – Anotator	P_o	P_e	κ	*Interpretasi*
ChatGPT* × *Annotator 1	**0.7241**	**0.6398**	**0.2341**	**Fair**
ChatGPT* × *Annotator 2	**0.7088**	**0.6169**	**0.2397**	**Fair**
ChatGPT* × *Annotator 3	**0.7154**	**0.6467**	**0.1944**	**Slight**
Gemini* × *Annotator 1	**0.8864**	**0.6458**	**0.6793**	**Substantial**
Gemini* × *Annotator 2	**0.8737**	**0.6154**	**0.6717**	**Substantial**
Gemini* × *Annotator 3	**0.8664**	**0.6350**	**0.6339**	**Substantial**
DeepSeek* × *Annotator 1	**0.9132**	**0.6320**	**0.7640**	**Substantial**
DeepSeek* × *Annotator 2	**0.9058**	**0.6057**	**0.7611**	**Substantial**
DeepSeek* × *Annotator 3	**0.8991**	**0.6206**	**0.7341**	**Substantial**

DeepSeek memperoleh nilai rata-rata Cohen's Kappa tertinggi (κ = 0,753), diikuti oleh Gemini (κ = 0,662) dan ChatGPT (κ = 0,223). Hasil tersebut menunjukkan bahwa DeepSeek memiliki tingkat kesepakatan tertinggi terhadap penilaian manusia, sedangkan ChatGPT menunjukkan tingkat kesesuaian yang relatif lebih rendah.

Selain itu, konsistensi hubungan antara hasil klasifikasi model dan *ground truth* dievaluasi menggunakan Spearman's Rank Correlation. Metode ini dipilih karena label sentimen bersifat ordinal, sehingga mampu mengukur hubungan monotonik tanpa mengasumsikan distribusi normal data. Sebelum perhitungan dilakukan, setiap label sentimen dikonversi ke dalam skala ordinal sebagaimana ditunjukkan pada Tabel 4.

Tabel 4. Kode Ordinal untuk Label Sentiment

Label Sentimen	Kode Ordinal
Negative	**1**
Neutral	**2**
Positive	**3**

Setelah proses konversi dilakukan, setiap data ulasan direpresentasikan sebagai sepasang nilai ordinal yang berasal dari hasil klasifikasi sentimen oleh model AI dan penilaian sentimen oleh annotator manusia. Secara matematis, pasangan data antara satu model AI dan satu annotator dinyatakan sebagai pasangan terurut (Xi,Yi) dengan i=1,2,…,n, dengan n menyatakan jumlah total ulasan yang dianalisis. Nilai Xi merepresentasikan skor ordinal sentimen yang dihasilkan oleh model AI pada ulasan ke-i, sedangkan nilai Yi merepresentasikan skor ordinal sentimen

yang diberikan oleh annotator pada ulasan ke-i yang sama. Pendekatan ini memastikan bahwa setiap pasangan (Xi ,Yi) membandingkan penilaian sentimen terhadap objek ulasan yang identik, sehingga memungkinkan pengukuran hubungan korelatif yang akurat menggunakan korelasi Spearman. Mengingat korelasi Spearman hanya melibatkan dua variabel, maka dalam penelitian ini setiap perhitungan dilakukan secara terpisah untuk satu model AI sebagai variabel X dan satu annotator sebagai variabel Y untuk setiap kombinasi yang dihitung secara terpisah.

Apabila suatu nilai muncul lebih dari satu kali (ties), maka peringkat yang digunakan adalah rata-rata dari seluruh posisi peringkat yang ditempati oleh nilai tersebut. Secara matematis, jika suatu nilai yang sama menempati peringkat ke-k hingga peringkat ke-m dan muncul sebanyak t=m−k+1 kali, maka peringkat yang diberikan dihitung sebagai

$$\text{Rank} = \frac{k + m}{2}$$

Dengan menggunakan rumus rata-rata peringkat, diperoleh:

- Rank (X1) $= \frac{1+67}{2} = 34$
- Rank (X2) $= \frac{68+327}{2} = 197{,}5$
- Rank (X3) $= \frac{68+327}{2} = 912{,}5$

dan hasil klasifikasi Annotator 1, diperoleh rentang peringkat sebagai berikut:

- Nilai 1 (Negatif) menempati peringkat 1 s.d. 143
- Nilai 2 (Netral) menempati peringkat 144 sampai 314
- Nilai 3 (Positif) menempati peringkat 314 sampai 1497

Dengan menggunakan rumus rata-rata peringkat, diperoleh:

- Rank (Y1) $= \frac{1+67}{2} = 72$
- Rank (Y2) $= \frac{68+327}{2} = 229$
- Rank (Y3) $= \frac{68+327}{2} = 906$

Setelah nilai peringkat untuk masing-masing kategori sentimen pada variabel X (model AI) dan variabel Y (Annotator) diperoleh, langkah selanjutnya adalah melakukan substitusi nilai peringkat tersebut ke setiap ulasan sesuai dengan nilai sentimen yang dimilikinya. Untuk X (ChatGPT) dan Y (Annotator 1) maka nilai substitusnya dapat dilihat pada tabel 5.

Tabel 5. Hasil Substitusi Nilai Peringkat Ke Setiap Ulasan

Nilai X	Rank X	Nilai Y	Rank Y
1	34	1	72
2	197,5	2	229
3	912,5	3	906

Dengan kata lain, setiap ulasan yang memiliki nilai sentimen tertentu akan memperoleh nilai peringkat yang sama berdasarkan hasil pemetaan peringkat yang telah ditentukan sebelumnya. Setelah pasangan peringkat untuk setiap ulasan diperoleh, dilakukan perhitungan **selisih peringkat** yang dinyatakan sebagai:

$$d_i = Rank(X_i) - Rank(Y_i)$$

Selanjutnya, nilai selisih peringkat tersebut dikuadratkan untuk menghilangkan pengaruh tanda negatif, sehingga diperoleh:

$$d_i{}^2 = (Rank(X_i) - Rank(Y_i))^2$$

Nilai dihitung untuk seluruh ulasan yang dianalisis dan kemudian disajikan dalam bentuk T**abel 6** ringkas yang selanjutnya digunakan dalam perhitungan koefisien korelasi Spearman pada Rumus (3).

Tabel 6. Perhitungan Koefisien Korelasi Spearman Rumus (3)

NO	X	Rank(X)	Y	Rank(Y)	d_i	$d_i{}^2$
1	3	912,5	3	906	6,5	42,25
2	3	912,5	1	72	840,5	706440,3
3	3	912,5	3	906	6,5	42,25
4	2	197,5	3	906	-708,5	501972,3
5	3	912,5	3	906	6,5	42,25
6	2	197,5	3	906	-708,5	501972,3
7	3	912,5	3	906	6,5	42,25

NO	X	Rank(X)	Y	Rank(Y)	d_i	d_i^2
8	3	912,5	3	906	6,5	42,25
9	3	912,5	2	229	683,5	467172,3
10	3	912,5	3	906	6,5	42,25
…	…	…	…	…	…	…
1497	3	912,5	3	906	6,5	42,25
∑di2						**181006708,5**

Maka untuk ChatGPT × Annotator 1 didapatkan

$$\rho = 1 - \frac{6\ x\ 181006708{,}5}{1479(1479^2 - 1)}$$

$$\rho = 0{,}3654$$

Perhitungan tersebut diterapkan secara konsisten pada seluruh kombinasi antara model AI dan annotator. Hasil perhitungan untuk seluruh kombinasi tersebut selanjutnya dirangkum dan disajikan pada Tabel 7 untuk memberikan gambaran komprehensif mengenai tingkat hubungan monotonic antara hasil klasifikasi sentimen oleh masing-masing model AI dan penilaian annotator manusia.

Tabel 7. Nilai Korelasi Spearman (ρ) antara Hasil Klasifikasi Sentimen Model LLM dan Anotator Manusia

Model AI	Annotator	ρ (Spearman)	Interpretasi
ChatGPT	Annotator 1	0,365	Lemah–Sedang
ChatGPT	Annotator 2	0,365	Lemah–Sedang
ChatGPT	Annotator 3	0,370	Lemah–Sedang
Gemini	Annotator 1	0,838	Sangat Kuat
Gemini	Annotator 2	0,864	Sangat Kuat
Gemini	Annotator 3	0,801	Kuat
DeepSeek	Annotator 1	0,866	Sangat Kuat
DeepSeek	Annotator 2	0,873	Sangat Kuat
DeepSeek	Annotator 3	0,841	Sangat Kuat

Sebagai analisis pendukung, dilakukan uji Chi-Square untuk membandingkan distribusi hasil klasifikasi model dengan distribusi *ground truth*. Seluruh kombinasi pengujian menghasilkan *p-value* < 0,001 (**Tabel 8**), yang menunjukkan bahwa distribusi prediksi masing-masing model berbeda secara signifikan dibandingkan distribusi anotasi manusia. Namun demikian, uji Chi-Square hanya menunjukkan adanya perbedaan distribusi dan tidak digunakan sebagai dasar penentuan performa model. performa model.

Tabel 8. Ringkasan Hasil Uji Chi-square (9 Kombinasi)

Model AI	Annotator	χ^2	df	p-value
ChatGPT	Annotator 1	326,720	4	< 0,001
ChatGPT	Annotator 2	331,045	4	< 0,001
ChatGPT	Annotator 3	301,881	4	< 0,001
Gemini	Annotator 1	1.203,161	4	< 0,001
Gemini	Annotator 2	1.308,999	4	< 0,001
Gemini	Annotator 3	1.091,480	4	< 0,001
DeepSeek	Annotator 1	1.613,121	4	< 0,001
DeepSeek	Annotator 2	1.626,276	4	< 0,001
DeepSeek	Annotator 3	1.485,570	4	< 0,001

Pembahasan

Hasil penelitian menunjukkan bahwa DeepSeek merupakan model yang memiliki tingkat kesesuaian paling tinggi terhadap *ground truth* manusia berdasarkan Cohen's Kappa maupun Spearman's Rank Correlation. Temuan ini mengindikasikan bahwa DeepSeek lebih konsisten dalam menginterpretasikan sentimen sesuai dengan persepsi manusia dibandingkan ChatGPT maupun Gemini. Selain menghasilkan tingkat kesepakatan yang lebih tinggi, distribusi prediksi DeepSeek juga paling mendekati distribusi *ground truth*, sehingga menunjukkan kemampuan yang lebih seimbang dalam membedakan sentimen positif, netral, dan negatif.

Gemini menunjukkan performa yang kompetitif dengan tingkat kesesuaian yang berada di bawah DeepSeek. Model ini cenderung lebih sensitif dalam mengidentifikasi sentimen negatif, sehingga menghasilkan jumlah prediksi negatif yang lebih tinggi dibandingkan distribusi *ground truth*. Karakteristik tersebut berpotensi memberikan keuntungan pada aplikasi yang berorientasi pada deteksi keluhan pelanggan, tetapi juga dapat meningkatkan kemungkinan klasifikasi negatif pada ulasan yang bersifat ambigu.

Sebaliknya, ChatGPT menunjukkan tingkat kesesuaian yang paling rendah terhadap *ground truth*. Distribusi hasil klasifikasi memperlihatkan kecenderungan model untuk memberikan label positif dalam proporsi yang lebih besar dibandingkan dua model lainnya. Pola tersebut mengindikasikan bahwa ChatGPT cenderung menginterpretasikan sebagian ulasan yang bersifat negatif atau ambigu sebagai sentimen positif maupun netral, sehingga menurunkan tingkat kesepakatan terhadap penilaian manusia.
emuan penelitian ini sejalan dengan berbagai penelitian sebelumnya yang menunjukkan bahwa performa *Large Language Models* (LLMs) dipengaruhi oleh domain aplikasi dan karakteristik data yang digunakan. Hasil penelitian juga mengindikasikan bahwa popularitas suatu model tidak selalu mencerminkan tingkat kesesuaiannya dengan penilaian manusia. Pemilihan model untuk analisis sentimen perlu didasarkan pada hasil evaluasi empiris pada domain yang relevan sehingga model yang dipilih benar-benar mampu menghasilkan klasifikasi yang lebih akurat dan konsisten dengan persepsi manusia, bukan semata-mata berdasarkan popularitas atau kemampuan generatifnya.

SIMPULAN

Penelitian ini mengevaluasi tingkat kesesuaian hasil analisis sentimen yang dihasilkan oleh ChatGPT, Gemini, dan DeepSeek terhadap ground truth yang dibentuk melalui anotasi manusia. Hasil pengujian menunjukkan bahwa ground truth memiliki reliabilitas yang baik dengan nilai Fleiss' Kappa sebesar 0,7053. Berdasarkan evaluasi menggunakan Cohen's Kappa dan Spearman's Rank Correlation, DeepSeek memperoleh tingkat kesesuaian tertinggi terhadap penilaian manusia, diikuti oleh Gemini, sedangkan ChatGPT menunjukkan tingkat kesesuaian yang paling rendah.

Hasil penelitian menunjukkan bahwa terdapat perbedaan kemampuan interpretasi sentimen di antara ketiga Large Language Models (LLMs). Temuan ini menegaskan bahwa pemilihan model LLM untuk analisis sentimen sebaiknya didasarkan pada hasil evaluasi empiris terhadap human ground truth, sehingga model yang digunakan mampu menghasilkan interpretasi yang lebih mendekati persepsi manusia.

Penelitian selanjutnya disarankan untuk memperluas evaluasi pada berbagai domain dan bahasa, serta melibatkan model LLM yang lebih baru agar diperoleh gambaran yang lebih komprehensif mengenai kemampuan interpretasi sentimen. Selain itu, penggunaan metrik evaluasi tambahan, seperti Matthews Correlation Coefficient (MCC) atau Cramér's V, dapat dipertim-

bangkan untuk melengkapi analisis performa model.

DAFTAR PUSTAKA

O. Campesato, Large Language Models: An Introduction. Germany: De Gruyter, 2024.

M. Ren, "Advancements and Applications of Large Language Models in Natural Language Processing: A Comprehensive Review," Applied and Computational Engineering, vol. 97, pp. 55–63, 2024. doi : 10.54254/2755-2721/97/20241406

T. R. Feyijimi, J. O. Aliu, A. E. Oke, and D. O. Aghimien, "ChatGPT's expanding horizons and transformative impact across domains: A critical review of capabilities, challenges, and future directions," Computers, vol. 14, no. 9, p. 366, 2025, doi: 10.3390/computers14090366

Z. Chkirbene, R. Hamila, A. Gouissem, and U. Devrim, "Large Language Models (LLM) in Industry: A Survey of Applications, Challenges, and Trends," in 2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET), Doha, Qatar, 2024, pp. 229-234, doi: 10.1109/HONET63146.2024.10822885.

H. Huang, A. A. Zavareh and M. B. Mustafa, "Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions," IEEE Access, vol. 11, pp. 90367-90382, 2023, doi: 10.1109/ACCESS.2023.3307308.

R. Sharma, B. Singh, and A. Khamparia, "Machine learning and generative AI techniques for sentiment analysis with applications," in Machine Learning and Generative AI Techniques for Sentiment Analysis with Applications, pp. 183–208, 2025. doi: 10.1002/9781394280735.ch10

S. Ivanov, K. Volchek, and C. Brito, "Generative AI for sentiment analysis," in Information and Communication Technologies in Tourism 2025. ENTER 2025, L. Nixon, A. Tuomi, and P. O'Connor, Eds. Cham: Springer, 2025, pp. 129 - 139. doi: 10.1007/978-3-031-83705-0_11

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, art. 248, pp. 1–38, Mar. 2023, doi: 10.1145/3571730.

J. Shuford, "Examining Ethical Aspects of AI: Addressing Bias and Equity in the Discipline," Journal of Artificial Intelligence General Science (JAIGS), vol. 3, no. 1, pp. 262–280, 2024, doi: 10.60087/jaigs.v3i1.119.

A. Jadon, "Ethical AI Development: Mitigating Bias in Generative Models," in Interplay of Artificial General Intelligence with Quantum Computing: Towards Sustainability, C. K. K. Reddy, S. Joseph, H. Joshi, M. Ouaissa, and M. M. Hanafiah, Eds. Cham: Springer Nature Switzerland, 2025, pp. 123–136, doi: 10.1007/978-3-031-87931-9_10.

F. Cabitza, A. Campagner, and V. Basile, "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing," in Thirty-Seventh AAAI Conf. Artif. Intell., Thirty-Fifth Conf. on Innovative Applications of AI, Thirteenth Symp. on Educational Advances in AI, B. Williams, Y. Chen, and J. Neville, Eds., Washington, DC, USA, Feb. 7–14, 2023, pp. 6860–6868, doi: 10.1609/aaai.v37i6.25840.

Aka, K. Burke, A. Bauerle, C. Greer, and M. Mitchell, "Measuring Model Biases in the Absence of Ground Truth," in AIES '21: Proc. 2021 AAAI/ACM Conf. AI, Ethics, and Society, Virtual Event, USA, pp. 327–335, 2021, doi: 10.1145/3461702.3462557.

S. Vanbelle, C. Engelhart, and E. Blix, "A comprehensive guide to study the agreement and reliability of multi-

observer ordinal data," BMC Medical Research Methodology, vol. 24, no. 310, 2024. [Online]. Available: https://doi.org/10.1186/s12874-024-02431-y.

W. Gan, Z. Qi, J. Wu, and J. C.-W. Lin, "Large language models in education: Vision and opportunities," in Proc. IEEE Int. Conf. Big Data (BigData), Sorrento, Italy, 2023, pp. 4776–4785. doi: 10.1109/BigData59044.2023.10386291

Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," Informatics, vol. 11, no. 3, p. 57, 2024. doi: 10.3390/informatics11030057

R. Singh et al., "ChatGPT vs. Gemini: Comparative accuracy and efficiency in Lung-RADS score assignment from radiology reports," Clinical Imaging, vol. 121, p. 110455, 2025. doi: 10.1016/j.clinimag.2025.110455

P. P. Nian et al., "ChatGPT and Google Gemini are Clinically Inadequate in Providing Recommendations on Management of Developmental Dysplasia of the Hip Compared to AAOS CPGs," Journal of the Pediatric Orthopaedic Society of North America, vol. 10, p. 100135, 2025. doi: 10.1016/j.jposna.2024.100135

Z. Zhan, S. Zhou, H. Zhou, J. Deng, Y. Hou, J. Yeung, and R. Zhang, "An evaluation of DeepSeek Models in Biomedical Natural Language Processing," arXiv preprint, arXiv:2503.00624, 2025. [Online]. Available: https://arxiv.org/abs/2503.00624

NadyinKy. Sephora products and skincare reviews [Internet]. Kaggle; 2023 [cited 2025 Mar 12]. Available from: https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews.

K. A. A. Al-Hameed, "Spearman's correlation coefficient in statistical analysis," Int. J. Nonlinear Anal. Appl., vol. 13, no. 1, pp. 3249–3255, 2022, doi: 10.22075/ijnaa.2022.6079.

Y. Sun, D. Sheng, Z. Zhou et al., "AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content," Humanities and Social Sciences Communications, vol. 11, art. 1278, 2024, doi: 10.1057/s41599-024-03811-x.

Permata, R., & Julianty, L. C. (2025). Towards an Automated Essay Evaluation System NLP Based Text Embeddings and Similarity Metrics. Digital Zone: Jurnal Teknologi Informasi dan Komunikasi, 16(1), 37–46. https://doi.org/10.31849/digitalzone.v16i1.26541.

D. C. Howell, "Chi-Square Test: Analysis of Contingency Tables," in International Encyclopedia of Statistical Science, M. Lovric, Ed. Berlin, Heidelberg: Springer, 2025. doi: 10.1007/978-3-662-69359-9_109.

S. Vanbelle, C. Engelhart, and E. Blix, "A comprehensive guide to study the agreement and reliability of multi-observer ordinal data," BMC Medical Research Methodology, vol. 24, no. 310, 2024. [Online]. Available: https://doi.org/10.1186/s12874-024-02431-y.