International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Sentiment Analysis Using Large Language Models: A Review

Simran A. Sodha[1], Prof. Neeta J. Sapara[2]

[1]PG Scholar, Department of Computer Science and Engineering, Dr. Subhash University, Junagadh, Gujarat, India
(ORCID: 0009-0003-2046-577X, Email: simran.sodha@dsuni.ac.in)

[2]Assistant Professor, Department of Computer Engineering, Dr. Subhash University, Junagadh, Gujarat, India
(ORCID: 0009-0006-4232-2624, Email: nita.sapara@dsuni.ac.in)

Abstract - Sentiment analysis is an important Natural Language Processing (NLP) task for identifying opinions, attitudes, and emotional polarity in user-generated text. Recent research has examined Large Language Models (LLMs) such as ChatGPT, Gemini, GPT-4, LLaMA, and related models for zero-shot, few-shot, prompt-based, multilingual, and aspect-based sentiment analysis. Compared with conventional supervised approaches, LLMs provide flexible natural-language task formulation, but published studies also report challenges related to consistency, domain adaptation, multilingual performance, fine-grained classification, sarcasm, explainability, and computational cost This paper presents a review of sentiment analysis using Large Language Models, with particular attention to user reviews and the analysis of generative-AI applications. The review covers traditional feature-based methods, LLM prompting and classification approaches, multilingual and aspect-based sentiment analysis, and application-oriented review mining. A five-class sentiment framework — very negative, negative, neutral, positive, and very positive — is considered as a fine-grained alternative to binary classification. The review further identifies research gaps involving binary-class limitations, dataset size and diversity, domain adaptation, and reliable evaluation. These observations motivate the proposed study of Google Gemini mobile-application reviews using TF-IDF and machine-learning classifiers, with evaluation through Accuracy, Precision, Recall, and F1-score.

Key Words: Sentiment Analysis, Large Language Models, Natural Language Processing, Fine-Grained Sentiment Analysis, Gemini, User Reviews.

1. INTRODUCTION

The rapid growth of digital platforms has created a large volume of user-generated textual information, including application reviews, product reviews, social-media posts, comments, and customer feedback. Such text contains information about user satisfaction, dissatisfaction, expectations, and experiences. Sentiment analysis provides a computational method for converting this unstructured information into structured sentiment categories and has therefore become an important application of Natural Language Processing [16].

Traditional sentiment analysis commonly combines text preprocessing, feature extraction, and supervised machine-learning algorithms. TF-IDF is one widely used representation for converting textual content into numerical features, after which classifiers can be trained for sentiment prediction. However, traditional approaches can be sensitive to vocabulary, domain, language, and feature-engineering choices [3], [16].
The emergence of Large Language Models has expanded the range of possible approaches. Research has compared LLMs with traditional sentiment classifiers and has investigated zero-shot and few-shot classification, prompt engineering, multilingual sentiment analysis, and aspect-based sentiment analysis [1], [10], [21]. Studies also show that LLM performance is not uniformly strong across all sentiment tasks, especially when the task requires fine-grained distinctions or robust handling of domain-specific language [10], [14].

A further issue is sentiment granularity. Binary classification generally separates text into positive and negative classes, but user reviews may contain neutral statements or different degrees of satisfaction and dissatisfaction. The MSR-1 research material therefore considers a five-class scale consisting of very negative, negative, neutral, positive, and very positive sentiment. This provides a more detailed representation of user feedback than a simple binary formulation.

Within this context, Google Gemini provides a relevant application domain for review-based sentiment analysis. The proposed research direction focuses on collecting Gemini mobile-application reviews and classifying them into five sentiment levels using TF-IDF and machine-learning techniques. The intended evaluation measures are Accuracy, Precision, Recall, and F1-score. The objective is not to assume that the proposed experiment has already produced results, but to establish a research framework from the findings of the reviewed literature[3], [16].

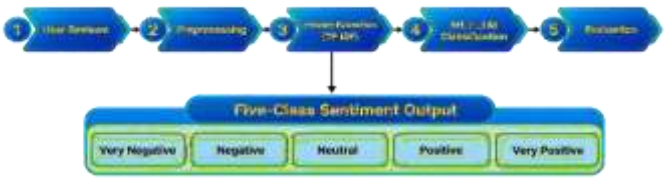


Fig. -1. Conceptual workflow for review-based fine-grained sentiment analysis.

2. FUNDAMENTALS OF SENTIMENT ANALYSIS

Sentiment analysis can be performed at document, sentence, or aspect level. At a basic level, a system assigns a polarity label to a text segment. More detailed systems attempt to identify the intensity of sentiment and the specific aspect toward which the sentiment is expressed[2]. Recent LLM literature has expanded these settings to multilingual and aspect-based classification [12], [18], [21].

A. Binary and Multi-Class Sentiment

Binary sentiment analysis generally distinguishes positive and negative opinions. Although this formulation is simple, it can lose information when a review is neutral or expresses only mild dissatisfaction. Multi-class sentiment analysis addresses this limitation by defining more than two categories.

B. Fine-Grained Five-Class Sentiment

The proposed research uses a five-level interpretation of star ratings: 1-star as very negative, 2-star as negative, 3-star as neutral, 4-star as positive, and 5-star as very positive. This five-class framework is intended to capture differences in sentiment intensity and provide a more detailed understanding of user opinions.



Fig -2. Five class sentiment analysis

3. TRADITIONAL SENTIMENT ANALYSIS TECHNIQUES

A conventional sentiment-analysis pipeline generally includes data collection, cleaning, tokenization, normalization, feature extraction, model training, and evaluation. Text preprocessing can remove noise and standardize the representation before classification[3].

TF-IDF assigns weights to terms according to their occurrence in a document and their relative rarity across the collection. It remains useful for supervised text classification because it provides a compact numerical representation that can be used with classifiers such as Logistic Regression, Support Vector Machine, Naive Bayes, Random Forest, Gradient Boosting, and XGBoost.s for sentiment classification[3], [16].

Traditional approaches can be computationally efficient and comparatively interpretable, but their performance can depend on feature design and the characteristics of the training data. Domain-specific vocabulary, multilingual content, sarcasm, and contextual meaning can make classification more difficult [5], [12], [16].

4. LARGE LANGUAGE MODELS FOR SENTIMENT ANALYSIS

Large Language Models are pretrained neural language models capable of processing and generating natural language. Their use for sentiment analysis has been studied through zero-shot, few-shot, prompt-based, fine-tuned, and hybrid approaches [18], [21], [24].

In zero-shot classification, the model receives an instruction and a review without task-specific labeled examples. Few-shot classification adds a small set of examples to the prompt. Prompt design can influence classification behavior and output format, while fine-tuning adapts a model to a particular task or domain[2], [24].

Research comparing LLMs with conventional sentiment systems indicates that LLMs can be useful across several settings, but they should not automatically be treated as universally superior sentiment classifiers. Benchmark studies have reported limitations in fine-grained classification, consistency, domain transfer, and multilingual or low-resource settings [11], [12], [23].

5. LITERATURE REVIEW

The reviewed studies cover multilingual sentiment analysis, multimodal sentiment analysis, social-media classification, aspect-based sentiment analysis, user-review mining, and comparisons between LLMs and traditional NLP or machine-learning methods[4], [6], [22].

Across the reviewed studies, LLM-based sentiment analysis shows a consistent trade-off between flexibility and reliability. Prompt-based models reduce the need for task-specific training data, while fine-tuned models generally provide stronger performance on structured and aspect-level tasks. Zhang et al. evaluated LLMs across 13 sentiment-analysis tasks and 26 datasets and reported that LLMs are competitive on simpler tasks but remain weaker on complex structured sentiment phenomena [10]. Krugmann and Hartmann further demonstrated the practical potential of generative AI for consumer sentiment analysis, while emphasizing the need for careful validation across domains [1].

Recent work increasingly focuses on aspect-based and low-resource settings. Šmíd et al. showed that fine-tuned LLaMA-based models can outperform prior methods on several ABSA tasks, whereas zero-shot and few-shot settings remain less consistent. Hellwig et al. used LLM-generated synthetic samples to address labelled-data scarcity in ABSA [23]. Fehle et al. extended this direction to resource-scarce English and German settings and found instruction fine-tuning effective across multiple ABSA subtasks [24]. Silcenco et al. contributed a retail corpus designed for ABSA with LLMs, reinforcing the importance of domain-specific, well-structured evaluation data Collectively, these findings support more fine-grained, domain-aware, multilingual and reproducible evaluation rather than relying only on binary sentiment labels[10], [12], [16].

Table -1: Literature Review

Year / Author	Study Title	Approach	Key Findings	Limitation / Gap
2024 Krugmann et al.	Sentiment Analysis in the Age of Generative AI [1]	LLM-based sentiment analysis	Evaluates generative-AI capability for consumer sentiment analysis	Domain validation and robustness remain important
2024 Šmíd et al.	LLaMA-Based Models for Aspect-Based Sentiment Analysis [2]	Fine-tuned LLaMA/Orca models	Fine-tuning performs strongly across multiple ABSA tasks	Zero-/few-shot settings lag behind fine-tuning
2025 Wojdecka et al.	Effectiveness of Large Language Models and Software Libraries in Sentiment Analysis [3]	Experimental comparison	Compare LLMs with traditional NLP tools	Explainability, cost, and domain adaptation remain challenges
2025 Jin, J.	Harnessing the Power of Large Language Models for Real-World Sentiment Classification on Social Media [4]	GPT/Gemini-based approaches	Study LLM sentiment classification in social media	Sarcasm, slang, and misinformation challenges
2025 Bojić, Ljubiša, et al.	Evaluating Large Language Models Against Human Annotators in Latent Content Analysis [5]	Gemini, GPT-4, LLaMA comparison	Compare LLM outputs with human annotation	Bias and inconsistent judgment
2024 Alabduljabbar, Reham	User-Centric AI: Evaluating the Usability of Generative AI Applications through User Reviews on App Stores [6]	VADER sentiment analysis	Analyze generative-AI app reviews	Limited to positive/negative sentiment
2025 Demirbilek et al.	Restaurant Review Sentiment and SWOT Analysis using AWS and GPT-4 Large Language Models [7]	ML + AWS Comprehend + GPT-4	Analyze restaurant reviews and SWOT factors	Domain-specific limitations
2024 Niimi et al.	Dynamic Sentiment Analysis with Local Large Language Models using Majority Voting [8]	Local LLM majority voting	Improve sentiment consistency with voting	Larger datasets and domains required
2023 Wang, Zengzhi, et al.	Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study [9]	Zero-shot/few-shot ChatGPT	Evaluate ChatGPT for sentiment analysis	Fine-tuned models can be stronger on some complex tasks
2024 Zhang, Wenxuan, et al.	Sentiment Analysis in the Era of Large Language Models: A Reality Check [10]	Multi-LLM benchmark	Benchmark multiple LLMs	Need realistic evaluation frameworks

Year / Author	Study Title	Approach	Key Findings	Limitation / Gap
2023 Simmering et al.	Large Language Models for Aspect-Based Sentiment Analysis [11]	Zero-shot/prompt-based methods	Study LLMs for ABSA	Fine-tuning and domain adaptation needed
2024 Wu, Chengyan, et al.	Zero-Shot Multilingual Aspect-Based Sentiment Analysis with Large Language Models [12]	Zero-shot LLM approach	Study multilingual ABSA	Low-resource languages remain challenging
2026 Ortiz Quispe et al.	Comparative Study of Chatbots for Sentiment Analysis [13]	ChatGPT, Gemini, Claude, LLaMA	Compare chatbot-based sentiment analysis	Standardized evaluation criteria needed
2023 Kocoń, Jan, et al.	ChatGPT as a Sentiment Analyzer: Evaluating Large Language Models on Sentiment Classification [14]	LLM + ML comparison	Compare ChatGPT and ML models	Domain and language variation
2023 Sun, Xiaofei, et al.	Large Language Models for Sentiment Analysis: A Survey [15]	Prompting, fine-tuning, instruction learning	Review LLM sentiment approaches	Unified evaluation and domain benchmarks needed
2023 Lu, Qiang, et al.	Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. [16]	Review of sentiment analysis methods	Highlights recent advances and challenges	Lacks standardized evaluation framework
2025 VarastehNezhad, Arya, et al.	AI in Mental Health: Emotional and Sentiment Analysis of Large Language Models' Responses [17]	Gemini Pro, GPT-4o, Claude	Analyze emotional/sentiment properties of LLM responses	Clinical validation and safety required
2025 Tandelon, Dennis, and Andry Chowanda	Exploring Prompt-Based Classification Using Large Language Models for Indonesian E-Commerce Sentiment Analysis [18]	Logistic Regression, XGBoost, Gradient Boosting + LLMs	Compare ML and prompt-based LLMs	Binary sentiment limitation
2024 Zhang, Hao, et al.	A Comprehensive Evaluation of ChatGPT for Text Classification Tasks [19]	Prompt engineering, zero-/few-shot	Evaluate ChatGPT for classification	Prompt and domain adaptation needed
2023 Zhao, Y., Zhang, H., & Wang, Y	Large Language Models Are Not Strong Sentiment Analyzers [20]	Zero-shot/few-shot LLMs	Examine LLM limitations in sentiment tasks	Fine-grained sentiment remains challenging
2024 Buscemi, Alessio, and Daniele Proverbio	ChatGPT vs Gemini vs LLaMA on Multilingual Sentiment Analysis [21]	Prompt-based sentiment classification	Compare LLMs for multilingual sentiment classification	Larger multilingual benchmark datasets required

Year / Author	Study Title	Approach	Key Findings	Limitation / Gap
2024 Yang, H. et al.	Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey [22]	Survey methodology	Review LLMs for multimodal sentiment analysis	More practical multimodal applications needed
2025 Hellwig, Nils Constantin et al.	Exploring LLMs for Synthetic Training Samples for ABSA [23]	GPT-3.5/Llama-generated training data	Synthetic data can improve low-resource ABSA	Quality control of generated labels/data is required
2026 Fehle, Jakob, et al.	Leveraging Fine-Tuning of LLMs for ABSA in Resource-Scarce Environments [24]	Instruction fine-tuning	Strong results in limited-data English/German ABSA	Efficiency, hyperparameters and generalization require care

6. RESEARCH GAP

The reviewed literature indicates that binary sentiment classification remains common, but binary labels may not adequately capture neutral opinions or different degrees of dissatisfaction[10], [16].

Fine-grained sentiment analysis is therefore an important research direction. A five-class framework can retain more information from user ratings and provide a clearer representation of sentiment intensity[6], [12], [23].

Dataset size and diversity are also recurring concerns. Several reviewed studies use domain-specific or localized datasets, indicating the value of larger and more representative collections for reliable evaluation[.6]

Domain adaptation, multilingual text, sarcasm, explainability, and consistency are additional challenges. These issues are particularly relevant when analyzing user reviews of generative-AI applications because the reviews may contain technical vocabulary, informal language, and mixed levels of satisfaction.

The combined evidence indicates a specific gap for app-review studies: many published evaluations emphasize binary or three-way polarity, benchmark datasets, or broad LLM comparisons, while fewer studies examine large-scale five-class sentiment intensity for reviews of generative-AI mobile applications. This gap motivates a structured evaluation of Google Gemini reviews in which star ratings can support very negative, negative, neutral, positive, and very positive classes, while model performance is assessed with consistent classification metrics[2],[10],[12].

7. PROPOSED RESEARCH DIRECTION

Based on the reviewed literature and the MSR-1 research objectives, the proposed study focuses on Google Gemini mobile-application reviews. Reviews can be collected and mapped to five sentiment categories: very negative, negative, neutral, positive, and very positive. After preprocessing, TF-IDF can be used for feature representation and traditional

machine-learning algorithms can be trained and compared. Performance can be measured using Accuracy, Precision, Recall, and F1-score. The proposed workflow is illustrated in Figure 1.

This review does not claim completed experimental results because the supplied MSR-1 material specifies the objectives and methodology but does not provide final model scores. The purpose of the proposed framework is to establish a reproducible direction for the subsequent research stage[10], [16].

8. FUTURE ENHANCEMENTS

The identified research gaps directly suggest several future enhancements. Larger and more diverse review collections should be used to improve representativeness and robustness, while multilingual and code-mixed reviews should be incorporated to reflect real app-store usage. Aspect-based sentiment analysis can extend overall polarity labels by identifying opinions about response quality, accuracy, speed, interface, features, and reliability[2], [11], [19]. Future models should also address sarcasm, implicit sentiment, slang, and context-dependent expressions, and explainability methods can be integrated to identify the words or phrases that drive predictions. A systematic comparison between TF-IDF-based machine-learning classifiers and prompt-based or fine-tuned LLMs such as Gemini, GPT, and LLaMA on common datasets would clarify accuracy–cost trade-offs. Finally, the framework can be extended to real-time review monitoring so that changes in user sentiment can be tracked over time. These enhancements respond directly to the observed limitations in granularity,

dataset diversity, domain adaptation, multilingual robustness, explain ability, and evaluation consistency[10], [12], [16].

9. CONCLUSION

This review examined sentiment analysis using Large Language Models and summarized research involving multilingual classification, aspect-based analysis, social-media sentiment, user reviews, and comparisons between LLMs and traditional machine-learning approaches. The literature shows that LLMs offer flexible prompt-based capabilities, but their performance depends on the task, domain, language, prompting strategy, and evaluation setting.

The review also highlights the limitation of binary sentiment analysis for applications where neutral opinions and subtle dissatisfaction are important. A five-class framework can provide a more detailed representation of user feedback. For the

proposed Google Gemini review-analysis study, TF-IDF and machine-learning classifiers provide a clear baseline, while Accuracy, Precision, Recall, and F1-score provide standard evaluation measures.

Future research can extend this work through larger datasets, multilingual and aspect-based analysis, LLM-versus-ML comparisons, explainability, and real-time monitoring. These directions can support the development of more systematic and reliable sentiment-analysis systems for generative-AI applications.

10. REFERENCES

[1] Krugmann, Jan Ole, and Jochen Hartmann. "Sentiment analysis in the age of generative AI." Customer Needs and Solutions 11.1 (2024): 3.

[2] Šmíd, Jakub, Pavel Přibáň, and Pavel Kral. "LLaMA-based models for aspect-based sentiment analysis." Proceedings of the 14th workshop on computational approaches to subjectivity, sentiment, & social media analysis. 2024.

[3] Wojdecka, Agnieszka, Jakub Gromadziński, and Krzysztof Walczak. "Effectiveness of large language models and software libraries in sentiment analysis." Applied Computer Science 21.2 (2025).

[4] Jin, J. "Harnessing the Power of Large Language Models for Real-World Sentiment Classification on Social Media." ADVANCES IN ECONOMICS, MANAGEMENT AND POLITICAL SCIENCES Учредители: Eliwise Academy Ltd 156.1 (2025): 214-224.

[5] Bojić, Ljubiša, et al. "Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm." Scientific reports 15.1 (2025): 11477.

[6] Alabduljabbar, Reham. "User-centric AI: evaluating the usability of generative AI applications through user reviews on app stores." PeerJ Computer Science 10 (2024): e2421.

[7] Demirbilek, Mustafa, and Sevim Özulukale Demirbilek. "Restaurant Review Sentiment and SWOT Analysis: using AWS and GPT-4 Large Language Models." Acta Infologica 9.2 (2025): 491-511.

[8] Niimi, Junichiro. "Dynamic sentiment analysis with local large language models using majority voting: A study on factors affecting restaurant evaluation." arXiv preprint arXiv:2407.13069 (2024).

[9] Wang, Zengzhi, et al. "Is ChatGPT a good sentiment analyzer? A preliminary study." arXiv preprint arXiv:2304.04339 (2023).

[10] Zhang, Wenxuan, et al. "Sentiment analysis in the era of large language models: A reality check." Findings of the association for computational linguistics: NAACL 2024. 2024.

[11] Simmering, Paul F., and Paavo Huoviala. "Large language models for aspect-based sentiment analysis." arXiv preprint arXiv:2310.18025 (2023).

[12] Wu, Chengyan, et al. "Evaluating zero-shot multilingual aspect-based sentiment analysis with large language models." International Journal of Machine Learning and Cybernetics 16.10 (2025): 8079-8101.

[13] Ortiz Quispe, Brayan Salomon, Andy Misael Tafur Gomez, and Emigdio Antonio Alfaro Paredes. "Comparative study of chatbots for sentiment analysis." Humanities and Social Sciences Communications (2026).

[14] Kocoń, Jan, et al. "ChatGPT: Jack of all trades, master of none." Information fusion 99 (2023): 101861.

[15] Sun, Xiaofei, et al. "Sentiment analysis through llm negotiations." arXiv preprint arXiv:2311.01876 (2023).

[16] Lu, Qiang, et al. "Sentiment analysis: Comprehensive reviews, recent advances, and open challenges." IEEE Transactions on Neural Networks and Learning Systems 35.11 (2023): 15092-15112.

[17] VarastehNezhad, Arya, et al. "AI in Mental Health: Emotional and Sentiment Analysis of Large Language Models' Responses to Depression, Anxiety, and Stress Queries." arXiv preprint arXiv:2508.11285 (2025).

[18] Tandelon, Dennis, and Andry Chowanda. "Exploring Prompt-Based Classification Using Large Language Models for Indonesian E-Commerce Sentiment Analysis." 2025 5th International Conference of Science and Information Technology in Smart Administration (ICSINTESA). IEEE, 2025.

[19] Zhang, Hao, et al. "Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: a comprehensive survey: H. Zhang et al." Artificial Intelligence Review 57.2 (2024): 17.

[20] Zhao, Y., Zhang, H., & Wang, Y. (2023). Large Language Models Are Not Strong Sentiment Analyzers.

[21] Buscemi, Alessio, and Daniele Proverbio. "Chatgpt vs gemini vs llama on multilingual sentiment analysis." arXiv preprint arXiv:2402.01715 (2024).

[22] Yang, H. et al. (2024). Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey. arXiv:2406.08068..

[23] Hellwig, Nils Constantin, Jakob Fehle, and Christian Wolff. "Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings." Expert Systems with Applications 261 (2025): 125514.

[24] Fehle, Jakob, et al. "Leveraging fine-tuning of large language models for aspect-based sentiment analysis in resource-scarce environments." Knowledge-Based Systems (2026): 115277.