Proceedings of the 4th International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/156/2025.20662

Harnessing the Power of Large Language Models for Real-World Sentiment Classification on Social Media

Jonathan Jin[1,a,*]

[1]*School of Literature, Science, and Art, University of Michigan Ann Arbor, MI, USA*
a. jonathan_jin@icloud.com
**corresponding author*

Abstract: Social media provides an abundant source of customer feedback, making sentiment analysis critical for businesses to understand consumer preferences and inform strategies. This study evaluates the effectiveness of Large Language Models (LLMs), such as GPT-3.5, GPT-4, and Google Gemini, in sentiment classification compared to traditional machine learning (ML) and deep learning (DL) models. Two benchmark datasets, the Stanford NLP/IMDB dataset and the FinanceInc/Auditor Sentiment dataset, are used to assess performance on general and domain-specific tasks. Advanced techniques, including zero-shot prompting, few-shot prompting, and fine-tuning, are applied to optimize LLMs. Results show that fine-tuned LLMs achieve the highest accuracy, outperforming ML and DL baselines. This study demonstrates the transformative potential of LLMs for sentiment analysis, offering scalable and efficient solutions for processing unstructured data. By addressing key limitations, LLMs enable deeper insights into customer sentiment, contributing to improved business intelligence and decision-making.

Keywords: Large Language Models, Artificial Classification, Business Applications
JEL codes: C83, C88, D83, L86, O33, G14

1. Introduction

The analysis of sentiment expressed in social media content has become a critical aspect of business intelligence, driven by the rapid growth of digital platforms as a primary source of customer feedback. Platforms such as Twitter, Facebook, and Instagram provide businesses with an unparalleled opportunity to understand customer preferences, opinions, and emerging trends [1]. However, the sheer volume of unstructured textual data presents a formidable challenge for manual processing, necessitating the adoption of automated techniques. The ability to extract meaningful insights from this vast amount of data not only informs product development and marketing strategies but also enhances customer service by identifying and addressing potential issues proactively [2].

Sentiment analysis, as a subset of natural language processing (NLP), has traditionally relied on methods like machine learning and deep learning to classify sentiment from text. Metrics such as review ratings and volumes have provided useful insights into customer perceptions, but sentiment extracted directly from textual feedback offers a more nuanced understanding. Kim [3] highlighted that online reviews analyzed through sentiment classification yield more effective metrics for understanding customer feedback than conventional aggregate measures like ratings. As businesses

DOI: 10.54254/2754-1169/156/2025.20662

increasingly rely on these insights to refine their strategies, advancements in technology are proving indispensable.

Large Language Models (LLMs) represent a significant leap forward in the field of sentiment analysis. These models, trained on vast datasets, have the ability to understand and generate human-like text, making them powerful tools for analyzing sentiment in unstructured social media data. LLMs have already demonstrated transformative potential across various applications, such as conversational AI, image and video generation, task automation, and data analysis. For businesses, the ability to leverage LLMs for sentiment analysis offers a means to quickly process and interpret vast amounts of customer feedback, ultimately aiding in strategic decision-making. By automating the classification of sentiment, LLMs allow businesses to gain deeper insights into customer satisfaction, gauge the success of products and services, and address emerging market trends.

Despite their promise, LLMs are not without limitations. One key issue is their tendency to produce generic outputs, particularly when faced with specialized domains or contexts for which they have not been adequately trained. This can lead to inaccuracies in sentiment classification, as the models may fail to capture the nuances of domain-specific language [2]. Another challenge is the "hallucination problem," where LLMs generate misleading or incorrect outputs due to a lack of contextual understanding. These limitations highlight the need for improved methodologies, such as advanced prompting techniques and the integration of external knowledge bases, to enhance the performance and reliability of LLMs in sentiment classification tasks.

This study aims to explore how LLMs can be optimized to perform sentiment classification on social media content. Specifically, we focus on two approaches: using advanced prompting techniques, such as zero-shot and few-shot prompting, to guide the models toward more accurate outputs, and employing Retrieval-Augmented Generation (RAG) techniques to integrate external knowledge bases, thereby mitigating the risks of hallucination and enhancing domain-specific performance. By addressing these challenges, we aim to unlock the full potential of LLMs as tools for analyzing social media sentiment, offering businesses actionable insights from the wealth of customer feedback available online.

The research question addressed in this paper centers on the efficacy of LLMs in classifying sentiment from social media content and their potential to overcome limitations through advanced methods. Given the growing integration of AI technologies into business practices, the findings of this study are expected to have significant implications for businesses seeking to harness the power of social media analytics. Furthermore, the results will contribute to the broader field of NLP and AI by advancing the understanding of how LLMs can be tailored to specific tasks.

The paper is structured as follows: Section 2 reviews the literature on sentiment analysis and the evolution of AI approaches in this domain, emphasizing the importance of sentiment classification for businesses. Section 3 introduces the datasets and their characteristics. Section 4 outlines the methodology, including the techniques employed for fine-tuning LLMs. Section 5 presents the results, comparing the performance of various approaches to sentiment classification. Finally, Section 6 discusses the practical implications, limitations, and avenues for future research.

Through this work, we aim to bridge the gap between the theoretical capabilities of LLMs and their practical applications in analyzing social media sentiment, contributing to both academic research and industry practice.

2. Related Work

2.1. Importance of Sentiment Classification on Social Media

Sentiment classification has emerged as a pivotal tool for businesses to understand customer perceptions, inform strategies, and respond to feedback. Kim [3] demonstrated that analyzing

DOI: 10.54254/2754-1169/156/2025.20662

sentiment from online reviews using natural language processing (NLP) provides a more nuanced understanding of customer feedback than traditional metrics like ratings or review volumes. Agarwal [2] emphasized how advanced analytics derived from sentiment analysis can drive business strategies and actionable insights, showcasing its role in shaping organizational decisions. Similarly, Mousavi et al. [1] found that sentiment analysis plays a key role in managing customer care on social media platforms, enabling firms to improve service effectiveness.

The business utility of sentiment classification is also reflected in its impact on consumer-facing industries. Gunarathne et al. [4] explored the role of customer sentiment in shaping differential treatment in the airline industry, finding that social media sentiment significantly influenced how complaints were prioritized. Kane [5] illustrated how KLM Royal Dutch Airlines successfully integrated social media sentiment analysis into its real-time customer support system, transforming it into a leader in digital customer service. Collectively, these studies underscore the transformative power of sentiment classification in improving customer engagement and informing business strategies.

2.2. AI Approaches to Sentiment Classification

The evolution of AI methods for sentiment classification has progressed from traditional machine learning techniques to advanced neural networks and transformer-based models. Early studies, such as Nasukawa and Yi [6], outlined the use of machine learning and NLP techniques to capture sentiment polarity in text, laying the groundwork for automated sentiment classification. Pang and Lee [7] provided an extensive survey of sentiment analysis methods, highlighting their broader applications in business intelligence, political analysis, and consumer behavior.

Subsequent studies expanded on these foundational techniques by applying deep learning and more complex models. For example, Geler et al. [8] employed machine learning-based sentiment analysis on food service reviews, demonstrating that text-based sentiment indicators could accurately predict customer satisfaction. Li et al. [9] introduced a sentiment analysis model combining Word2Vec, Bi-GRU, and attention mechanisms for restaurant reviews, showing improved classification performance compared to traditional methods. These advancements illustrate the increasing sophistication of sentiment classification approaches, moving toward models capable of handling more complex textual data.

2.3. Advances in LLM for Data Analysis

The emergence of large language models (LLMs) represents a paradigm shift in sentiment classification and broader NLP tasks. LLMs, such as GPT-3 and BERT, are pre-trained on vast datasets, enabling them to generate and interpret human-like text. Recent studies have explored their potential to analyze sentiment at scale. For instance, Liu and Lopez [10] demonstrated how social media conversations could shape brand perceptions, with LLMs playing a pivotal role in identifying sentiment trends across platforms. Yang et al. [11] analyzed user-generated content on Facebook business pages, emphasizing the role of LLMs in understanding customer engagement through sentiment classification.

However, the integration of LLMs into sentiment analysis is not without challenges. Research has highlighted their limitations, such as the generic nature of outputs when faced with domain-specific tasks [2] and the hallucination problem, where incorrect or misleading responses may arise due to insufficient contextual understanding. To address these issues, advanced prompting techniques and Retrieval-Augmented Generation (RAG) methods have been proposed. These approaches link LLM outputs to external knowledge bases, enhancing their ability to provide accurate and contextually relevant sentiment analysis.

DOI: 10.54254/2754-1169/156/2025.20662

2.4. Business Applications of Sentiment Analysis

Beyond technical advancements, sentiment classification has shown significant applications across industries. Chevalier and Mayzlin [12] analyzed how sentiment in online book reviews influenced sales, demonstrating the economic impact of sentiment polarity. Similarly, Xu et al. [13] explored how online physician reviews affected patient demand, highlighting the role of sentiment in shaping consumer decision-making. In the hospitality industry, Mankad et al. [14] utilized automated text analysis to extract insights from hotel reviews, offering valuable feedback for improving service quality.

The financial sector also benefits from sentiment analysis. Gric et al. [15] conducted a meta-analysis showing how investor sentiment impacts stock returns, particularly in U.S. markets. Their findings underscore the broader applicability of sentiment classification across domains, from hospitality to finance, further emphasizing its importance for businesses and researchers alike.

2.5. Gaps in Existing Research

While prior studies have demonstrated the potential of sentiment classification, gaps remain in addressing the limitations of LLMs for this task. Current research often lacks detailed analyses of advanced techniques, such as structured prompting or RAG, which are crucial for improving the accuracy and reliability of LLM-based sentiment analysis. Additionally, the practical implications of using LLMs for specific business applications, particularly in social media sentiment classification, remain underexplored. By addressing these gaps, this study aims to contribute to the evolving field of sentiment analysis, providing actionable insights for both academic research and industry practice.

3. Data

The efficacy of large language models (LLMs) in sentiment classification heavily depends on the quality and diversity of the datasets used for evaluation and fine-tuning. For this study, two benchmark datasets are utilized to ensure robust experimentation and comparison: the **Stanford NLP/IMDB dataset** and the **FinanceInc/Auditor Sentiment dataset**. These datasets provide varied contexts for sentiment analysis, ranging from user-generated reviews in entertainment to domain-specific sentiment in financial news.

3.1. Stanford NLP/IMDB Dataset

The Stanford NLP/IMDB dataset is a widely recognized benchmark in sentiment analysis, consisting of 50,000 unique movie reviews. The dataset is evenly divided into two subsets: 25,000 reviews for training and 25,000 for testing. Each review is labeled with a binary sentiment, indicating whether the sentiment is positive or negative.

In addition to pre-processed data with vocabulary and text labels, the dataset provides unfiltered text for further experimentation. This flexibility allows researchers to test various preprocessing techniques, making it a versatile resource for sentiment analysis models. The IMDB dataset's simplicity and balance in sentiment classes make it an ideal starting point for evaluating the general sentiment classification performance of LLMs, particularly in zero-shot and few-shot scenarios.

3.2. FinanceInc/Auditor Sentiment Dataset

The FinanceInc/Auditor Sentiment dataset focuses on sentiment analysis in the domain of financial news, offering a domain-specific challenge for LLMs. This dataset comprises several thousand sentences extracted from English-language financial news articles. Each sentence is categorized by

DOI: 10.54254/2754-1169/156/2025.20662

sentiment, with phrases assigned numerical values to represent their positive, negative, or neutral sentiment correspondence.

The dataset's specificity to the financial domain provides an excellent opportunity to test the adaptability of LLMs in specialized contexts. Unlike the IMDB dataset, which deals with general user-generated content, the FinanceInc/Auditor Sentiment dataset requires the model to interpret sentiment within a professional and technical lexicon. This distinction allows for evaluating whether LLMs can overcome limitations, such as producing generic outputs or struggling with domain-specific terminology.

3.3. Relevance of the Datasets

The combination of the IMDB and FinanceInc/Auditor datasets ensures a comprehensive evaluation of the LLMs' capabilities across different types of content. While the IMDB dataset provides insights into the models' performance on general sentiment classification, the FinanceInc dataset tests their ability to handle domain-specific language and sentiment nuances. Together, these datasets facilitate a balanced analysis of the models' strengths and weaknesses, enabling a deeper understanding of their potential applications in real-world sentiment analysis tasks.

In subsequent sections, the experimental results derived from these datasets will be presented to assess the effectiveness of various prompting techniques, fine-tuning strategies in improving sentiment classification accuracy.

4. Methods

This section outlines the methodological framework adopted in the study, including the models and techniques employed to evaluate sentiment classification on social media data. The methods range from traditional machine learning approaches to advanced large language models (LLMs), ensuring a comprehensive exploration of techniques across different paradigms.

4.1. Machine Learning and Deep Learning Models

To establish baseline comparisons, traditional machine learning (ML) and deep learning (DL) models are utilized. These approaches provide insights into the performance of earlier sentiment classification techniques and serve as benchmarks for evaluating the efficacy of LLMs.

1. **Vectorization + Machine Learning Models**: Textual data is preprocessed using vectorization methods such as CountVectorizer and TF-IDF Vectorizer to convert unstructured text into numerical representations. These representations are then fed into machine learning classifiers like Random Forest and XGBoost, which are widely used for sentiment classification tasks due to their robustness and interpretability [7,12].
2. **Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)**: RNNs and their variant, LSTM, are employed to capture sequential dependencies in text. These models are particularly effective in processing textual data with contextual relationships [8]. However, their performance is often limited by vanishing gradient issues and difficulty in handling long-range dependencies.
3. **Transformer-based Models (BERT)**: Bidirectional Encoder Representations from Transformers (BERT) represents a significant advancement in NLP. BERT pre-trains a deep bidirectional representation, enabling it to understand the context of a word based on both preceding and succeeding words. This ability makes BERT a strong contender for sentiment classification tasks, as it captures the subtle nuances of language more effectively than earlier approaches.

DOI: 10.54254/2754-1169/156/2025.20662

4.2. Large Language Models (LLMs)

Large Language Models (LLMs), such as GPT-3.5, GPT-4, and Google Gemini, represent the state of the art in natural language processing (NLP). These models leverage vast pretraining datasets and advanced architectures to generate and interpret human-like text. Their flexibility and capability to perform tasks without task-specific fine-tuning make them pivotal for applications such as sentiment analysis. In this study, the Gemini-1.5 model was evaluated and optimized using several approaches, including zero-shot prompting, few-shot prompting, and fine-tuning.

1. In **zero-shot prompting**, the model is queried directly without providing task-specific examples. A natural language prompt describes the task, such as "Classify the sentiment of this review as positive or negative." This approach relies on the model's pretraining knowledge to generalize and produce accurate outputs. For evaluation, 100 reviews were randomly sampled, and the task was performed 10 times. The average performance was reported, offering a measure of the model's ability to handle general sentiment analysis without prior task-specific context. Zero-shot prompting is quick and straightforward but may lack precision for nuanced or domain-specific tasks. The model's output quality heavily depends on prompt clarity and structure, introducing variability in results. Despite these challenges, zero-shot prompting is a useful baseline for tasks where labeled data is unavailable or when a rapid evaluation of the model's capabilities is required.
2. **Few-shot prompting** provides the model with a small number of labeled examples within the prompt, offering contextual clues about the task. For this study, five labeled reviews (randomly sampled) were included in each prompt, followed by 100 randomly sampled reviews for classification. This process was repeated 10 times, and the average performance was reported. For instance, the prompt might include examples such as: i) "Review: This product is fantastic! I love it. Sentiment: Positive"; ii) "Review: I didn't like the quality of this item. Sentiment: Negative" Few-shot prompting improves accuracy by leveraging these labeled examples, helping the model better understand task requirements. However, this approach faces input length constraints, as including examples reduces the space available for the task input. Performance also depends on the representativeness of the examples provided. While few-shot prompting generally outperforms zero-shot prompting, it still lacks the specificity and robustness achieved through fine-tuning.
3. **Fine-tuning** represents a more advanced approach to optimize LLMs for specific tasks. In this study, the Gemini-1.5 model was fine-tuned using 500 reviews, balanced with 250 positive and 250 negative examples. This task-specific training adjusted the model's weights to better capture the nuances of sentiment classification. For evaluation, 100 randomly sampled reviews were classified 10 times, and the average performance was reported. Fine-tuning significantly improves accuracy and allows the model to excel in domain-specific applications. However, it requires high-quality labeled data and computational resources, which can be limiting for some use cases. Moreover, fine-tuning narrows the model's focus, potentially reducing its flexibility for general tasks.

DOI: 10.54254/2754-1169/156/2025.20662

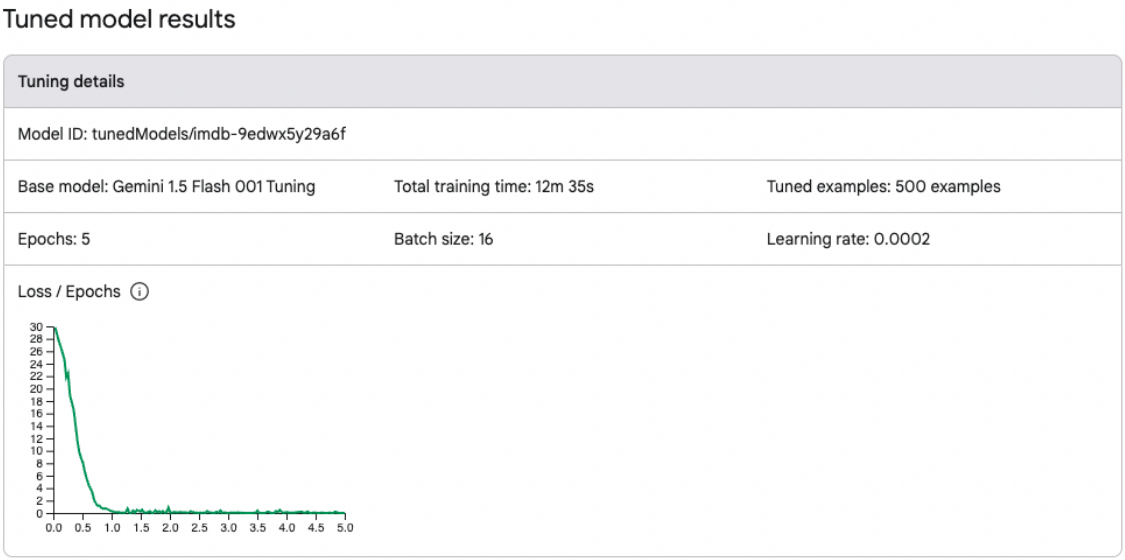

Figure 1: Fine-tuned Gemini model using 500 training examples (250 positive and 250 negative)

In summary, each of these approaches—zero-shot prompting, few-shot prompting and fine-tuning—has distinct advantages and limitations. Zero-shot prompting is quick and easy to apply but may lack precision. Few-shot prompting offers improved accuracy with minimal effort but depends heavily on the quality of examples. Finally, fine-tuning delivers high task-specific accuracy but demands significant data and computational resources, making it less flexible for general use. The choice of approach depends on the task requirements, availability of labeled data, and computational constraints, making it essential to balance these factors when deploying LLMs for NLP applications.

4.3. Experimental Setup

Each model is evaluated on two benchmark datasets: the Stanford NLP/IMDB dataset and the FinanceInc/Auditor Sentiment dataset. The ML models rely on traditional vectorization methods, while DL models use tokenized representations of text. For LLMs, different prompting strategies and fine-tuning techniques are applied. Metrics such as F1-score are used to compare performance across models and methods.

This methodological framework highlights the relative strengths and limitations of each technique. Zero-shot prompting serves as a baseline, few-shot prompting offers incremental improvements with minimal labeled data, and fine-tuning maximizes accuracy for specialized tasks. The findings highlight the value of leveraging advanced prompting and fine-tuning strategies to optimize LLMs like Gemini-1.5 for sentiment analysis, contributing both to academic research and practical applications. By systematically exploring these techniques, this study provides actionable insights into deploying LLMs effectively in real-world scenarios.

5. Results

The experimental results provide a detailed comparative analysis of sentiment classification performance across traditional machine learning (ML), deep learning (DL), and large language model (LLM) approaches. The findings reveal a clear trajectory of improvement, with advanced models like LLMs significantly outperforming traditional techniques, particularly when fine-tuned for specific datasets.

DOI: 10.54254/2754-1169/156/2025.20662

Table 1: Comparisons of performance

Model	IMDB	Auditor
Countvectorizer + Random Forest	85%	75%
TfidfVectorizer + Random Forest	84%	73%
Countvectorizer + XGBoost	85%	78%
TfidfVectorizer + XGBoost	86%	77%
RNN	67%	58%
LSTM	67%	58%
BERT	92.3%	86.1%
Roberta	86.8%	87.2%
LLM (zero-shot)	85%	80%
LLM (few-shot)	87.3%	82.5%
LLM fine-tuned	94.5%	89.4%

Traditional ML models performed reasonably well, with XGBoost showing a slight edge over Random Forest. Using vectorization methods such as CountVectorizer and TF-IDF Vectorizer, these models achieved accuracy rates ranging from 73% to 86% across both datasets. XGBoost with TF-IDF achieved the highest accuracy among ML models, at 86% for the IMDB dataset and 78% for the Auditor Sentiment dataset.

Deep learning models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory networks (LSTM) struggled, with accuracies hovering around 58% for the Auditor dataset and 67% for the IMDB dataset. These results reflect the inherent limitations of RNN-based models in handling long-range dependencies and capturing nuanced sentiment features.

Transformer-based models demonstrated a significant leap in performance. BERT achieved 92.3% accuracy on the IMDB dataset and 86.1% on the Auditor dataset, highlighting its superior ability to contextualize and process text. Similarly, RoBERTa reached 87.2% accuracy on the Auditor dataset, slightly outperforming BERT in domain-specific contexts.

5.1. LLM Performance and the Impact of Fine-Tuning

LLMs outperformed both ML and DL models across all experiments. The zero-shot prompting approach yielded solid results, with accuracies of 85% for the IMDB dataset and 80% for the Auditor dataset. Few-shot prompting further improved performance, achieving 87.3% on IMDB and 82.5% on Auditor reviews. These results show that even without task-specific training, LLMs exhibit strong generalization capabilities when contextual cues are provided.

The fine-tuned Gemini-1.5 model delivered the highest accuracy across all approaches, achieving 94.5% on the IMDB dataset and 89.4% on the Auditor dataset. The substantial improvement over zero-shot and few-shot prompting indicates the effectiveness of fine-tuning in adapting LLMs to specific datasets. By training on 500 balanced examples (250 positive and 250 negative), the fine-tuned model learned to recognize nuanced patterns and domain-specific sentiment expressions more effectively than other methods.

In particular, the fine-tuned LLM exhibited a stronger performance on the IMDB dataset compared to the Auditor dataset. This disparity can be attributed to differences in the nature of the datasets. The IMDB dataset, composed of user-generated movie reviews, features less technical language and simpler sentiment cues. This aligns well with the pretrained knowledge of LLMs and their ability to process general text.

In contrast, the Auditor dataset involves financial news, characterized by domain-specific terminology and complex sentiment expressions. While the fine-tuned LLM adapted well to these

DOI: 10.54254/2754-1169/156/2025.20662

challenges, achieving 89.4% accuracy, the domain's complexity presented a higher barrier than the more straightforward language in the IMDB dataset. This highlights the importance of fine-tuning when applying LLMs to specialized contexts.

6. Discussions and Conclusions

The results of this study indicate the transformative potential of large language models (LLMs) in sentiment classification, particularly when compared to traditional machine learning (ML) and deep learning (DL) models. By leveraging advanced techniques such as fine-tuning and structured prompting, LLMs not only surpass traditional approaches in accuracy but also exhibit adaptability to both general and domain-specific tasks. The findings provide several insights into the practical implications, limitations, and future directions for research and application.

Moreover, this study highlights the significant advancements in sentiment classification achieved through LLMs, particularly when combined with fine-tuning and advanced prompting techniques. The superior performance of LLMs across both general and domain-specific datasets demonstrates their versatility and potential for real-world applications. By addressing the limitations and exploring future research directions, LLMs can be further optimized to provide more accurate, reliable, and ethical sentiment analysis solutions.

The findings contribute to the broader field of natural language processing (NLP) by advancing the understanding of how LLMs can be tailored to specific tasks. They also offer valuable insights for businesses seeking to harness the power of sentiment analysis in understanding customer feedback, shaping strategies, and improving decision-making. As LLM technology continues to evolve, its integration into sentiment analysis workflows promises to transform how businesses and researchers analyze and interpret human sentiment on a global scale.

6.1. Practical implications

The ability of LLMs to analyze sentiment at scale presents significant opportunities for businesses across industries. Fine-tuned LLMs achieved the highest performance in both datasets, demonstrating their capability to handle domain-specific language and context. For example, in the financial domain, fine-tuned LLMs can accurately interpret sentiment from complex and technical texts, enabling financial institutions to better assess market sentiment or customer perceptions. Similarly, in consumer-oriented industries like entertainment and e-commerce, LLMs can provide more nuanced insights into customer feedback, allowing businesses to refine product offerings, marketing strategies, and customer service initiatives.

The integration of advanced prompting techniques, such as few-shot prompting further enhances the utility of LLMs for sentiment analysis tasks. These techniques enable businesses to deploy LLMs effectively in scenarios where labeled data is sparse or unavailable, such as analyzing feedback from emerging markets or new product categories. Moreover, the automation of sentiment classification through LLMs reduces the time and cost associated with manual analysis, making it accessible to a broader range of organizations.

6.2. Limitation

While LLMs demonstrate superior performance, several limitations must be acknowledged. One key challenge is their dependence on high-quality datasets for fine-tuning. The accuracy of fine-tuned LLMs heavily relies on the availability of well-labeled and representative data, which may not always be feasible in specialized or niche domains. In addition, fine-tuning requires significant computational resources, making it less accessible for small or resource-constrained organizations.

DOI: 10.54254/2754-1169/156/2025.20662

Another limitation is the susceptibility of LLMs to produce generic outputs or hallucinations when faced with ambiguous or poorly contextualized inputs. While advanced prompting techniques can mitigate this issue to some extent, they do not entirely eliminate the risk of inaccurate predictions. Furthermore, the reliance on carefully designed prompts in zero-shot and few-shot approaches may limit their scalability in real-world applications, where inputs can vary widely in structure and quality.

Ethical considerations also arise with the use of LLMs, particularly in sensitive domains like finance or healthcare. The opacity of LLM decision-making processes can make it difficult to identify biases or errors, potentially leading to unintended consequences if the models are deployed without adequate oversight.

6.3. Future Work

To address these limitations and further enhance the applicability of LLMs in sentiment analysis, future research should explore several avenues. First, future research can explore automated approaches to optimize prompt design for zero-shot and few-shot prompting. This would reduce the reliance on manual trial-and-error and enable more scalable deployment of LLMs in real-world scenarios. Second, combining LLMs with traditional or domain-specific models can yield improved performance, particularly in specialized applications. For example, integrating domain-specific ontologies with LLMs could enhance their ability to interpret technical language. Third, improving the interpretability of LLMs is crucial for building trust and ensuring ethical use. Techniques such as explainable AI (XAI) can be integrated into LLM frameworks to provide transparency into model predictions and decision-making processes.

References

[1] *Mousavi, R., Johar, M., & Mookerjee, V. S. (2020). The voice of the customer: Managing customer care in Twitter. Information Systems Research, 31(2), 340-360.*

[2] *Agarwal, S. (2022). Deep learning-based sentiment analysis: Establishing customer dimension as the lifeblood of business management. Global Business Review, 23(1), 119-136.*

[3] *Kim, R. Y. (2021). Using online reviews for customer sentiment analysis. IEEE Engineering Management Review, 49(4), 162-168.*

[4] *Gunarathne, P., Rui, H., & Seidmann, A. (2018). When Social Media Delivers Customer Service. MIS Quarterly, 42(2), 489-520.*

[5] *Kane, G. C. (2014). How Facebook and Twitter are reimagining the future of customer service. MIT Sloan Management Review, 55(4), 1.*

[6] *Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77).*

[7] *Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.*

[8] *Geler, Z., Savić, M., Bratić, B., Kurbalija, V., Ivanović, M., & Dai, W. (2021). Sentiment prediction based on analysis of customers assessments in food serving businesses. Connection Science, 33(3), 674-692.*

[9] *Li, L., Yang, L., & Zeng, Y. (2021). Improving sentiment classification of restaurant reviews with attention-based bi-GRU neural network. Symmetry, 13(8), 1517.*

[10] *Liu, Y., & Lopez, R. A. (2016). The impact of social media conversations on consumer brand choices. Marketing Letters, 27(1), 1–13.*

[11] *Yang, M., Ren, Y., & Adomavicius, G. (2019). Understanding user-generated content and customer engagement on Facebook business pages. Information Systems Research, 30(3), 839-855.*

[12] *Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345-354.*

[13] *Xu, Y., Armony, M., & Ghose, A. (2021). The interplay between online reviews and physician demand: An empirical investigation. Management Science, 67(12), 7344-7361.*

[14] *Mankad, S., Han, H. S., Goh, J., & Gavirneni, S. (2016). Understanding online hotel reviews through automated text analysis. Service Science, 8(2), 124-138.*

DOI: 10.54254/2754-1169/156/2025.20662

[15] Gric, Z., Bajzík, J., & Badura, O. (2023). Does sentiment affect stock returns? A meta-analysis across survey-based measures. International Review of Financial Analysis, 89, 102773.