Customer Needs and Solutions (2024) 11:3
https://doi.org/10.1007/s40547-024-00143-4

RESEARCH ARTICLE

Sentiment Analysis in the Age of Generative AI

Jan Ole Krugmann[1] · Jochen Hartmann[1]

Accepted: 13 February 2024 / Published online: 5 March 2024

Abstract

In the rapidly advancing age of Generative AI, Large Language Models (LLMs) such as ChatGPT stand at the forefront of disrupting marketing practice and research. This paper presents a comprehensive exploration of LLMs' proficiency in sentiment analysis, a core task in marketing research for understanding consumer emotions, opinions, and perceptions. We benchmark the performance of three state-of-the-art LLMs, i.e., GPT-3.5, GPT-4, and Llama 2, against established, high-performing transfer learning models. Despite their zero-shot nature, our research reveals that LLMs can not only compete with but in some cases also surpass traditional transfer learning methods in terms of sentiment classification accuracy. We investigate the influence of textual data characteristics and analytical procedures on classification accuracy, shedding light on how data origin, text complexity, and prompting techniques impact LLM performance. We find that linguistic features such as the presence of lengthy, content-laden words improve classification performance, while other features such as single-sentence reviews and less structured social media text documents reduce performance. Further, we explore the explainability of sentiment classifications generated by LLMs. The findings indicate that LLMs, especially Llama 2, offer remarkable classification explanations, highlighting their advanced human-like reasoning capabilities. Collectively, this paper enriches the current understanding of sentiment analysis, providing valuable insights and guidance for the selection of suitable methods by marketing researchers and practitioners in the age of Generative AI.

Keywords Generative AI · Large language models · Sentiment analysis · Machine learning · Digital marketing

1 Introduction

The recent emergence and rapid adoption of Large Language Models (LLMs) are disrupting the marketing landscape. McKinsey & Company's survey on the state of Generative AI indicates that the marketing and sales functions are the primary adopters of Generative AI tools [14]. Pioneering academic papers underscore the *dual role of Generative AI* in marketing. First, Generative AI is used in content creation, including creative writing tasks [47, 55], conversational customer support [9], or market research based on emulated consumers [7, 40]. For instance, Reisenbichler et al. (2022) demonstrate that natural language generation for search engine optimization (SEO) outperforms content created by human writers in search engine rankings, increasing overall campaign performance while reducing production costs [55]. Noy and Zhang (2023) report productivity and quality increases through the use of Generative AI tools for professional writing tasks [47]. Brynjolfsson et al. (2023) report a 14% productivity increase using Generative AI-based conversational assistants for customer service representatives [9]. Second, Generative AI is pioneered in zero-shot content analysis, which includes areas like visual analysis [38] and automated textual analysis [36, 41, 54, 66]. For instance, Konrad and Hartmann (2023) explore the versatility of multi-modal LLMs for visual content analysis [38]. Relatedly, Rathje et al. (2023) explore GPT's capabilities for multilingual psychological text analysis [54].

The role of automated text analysis in marketing, underscored by Berger et al.(2020) [2], is set to further expand with the adoption of Generative AI, as it is expected to boost not only the accuracy but also the accessibility of text mining methods. Sentiment analysis represents one of the most prevalent use cases of automated text analysis in marketing, offering deep insights into consumer emotions,

✉ Jan Ole Krugmann
jan.krugmann@tum.de

Jochen Hartmann
jochen.hartmann@tum.de

[1] Technical University of Munich (TUM), TUM School of Management, Arcisstr. 21, 80333 Munich, Germany

opinions, and perceptions [26, 30]. Applications of sentiment analysis include generating market insights from user-generated online content [45], predicting virality from linguistic features of newspaper articles [3], or identifying top performing individuals by analyzing the language style used in emails [67]. The field of sentiment analysis employs diverse methods, ranging from lexicon-based approaches to traditional machine learning models, and extending to the more advanced transfer learning techniques, as detailed by Hartmann et al. (2019 & 2023) [24, 25]. Demonstrating a significant advancement, Hartmann et al. (2023) find a superior performance of transfer learning models in sentiment analysis tasks compared to lexicon-based and traditional machine learning approaches, by 20 and 10 percentage points in accuracy, respectively [24].

Generative AI, especially LLMs, which are an advanced form of transfer learning models, show promise in further transforming sentiment analysis. The considerable scale of data used for LLM training could significantly increase the performance across sentiment analysis tasks and thus influence the choice of methods in this domain. Unlike the task-specific fine-tuning required for supervised problems with transfer learning, LLMs operate through natural language prompts, offering increased versatility across a broader range of applications. Instead of collecting labelled training data and fine-tuning a supervised machine learning model, users simply need to instruct the model what features they want to extract from a text, e.g., "*Classify the sentiment in these reviews. Only use positive or negative as sentiment scale*". This adaptability not only enhances the accessibility of LLMs but also makes them suitable for various sentiment classification tasks, from binary to multi-class [36, 66], and enables their application in zero-shot or few-shot scenarios [61].

Despite these advancements, there is a notable gap in comprehensive benchmarking of LLMs against established transfer learning models, particularly in the investigation of factors influencing classification accuracy such as data origin and textual data characteristics. Our paper addresses this research gap, extending the empirical framework of Hartmann et al. (2023) by incorporating recent general-purpose LLMs, thereby offering a refined guide for method selection in sentiment analysis in the age of Generative AI. Specifically, we conduct three experiments to benchmark the capabilities of three state-of-the-art LLMs, namely GPT-3.5, GPT-4 and Llama 2, in different sentiment analysis tasks against best-in-class transfer learning models.

First, we conduct a comparative study of LLMs in binary and three-class sentiment classification on over 3,900 unique text documents from 20 different datasets in a zero-shot setting. Zero-shot inference refers to a model's ability to execute tasks without previous domain-specific training [39]. We find that state-of-the-art LLMs are on par or even outperform traditional transfer learning methods in binary and three-class sentiment classification tasks.

Second, we investigate how textual data characteristics, including dataset origin, sentence and word count, and additional linguistic characteristics such as the presence of lengthy words, relate to classification accuracy. Additionally, we assess the effects of variations in the analytical procedure, particularly in terms of the prompting method and the provision of examples within the prompt, distinguishing between zero-shot and few-shot settings. We observe that text documents originating from Twitter[1] (as opposed to less noisy consumer reviews), along with the inclusion of a third output class (i.e., "neutral" in three-class sentiment analysis), and documents comprising only a single sentence (vs. multiple) tend to reduce classification accuracy across all tested LLMs. Conversely, the presence of lengthy words in text documents increases classification performance for all LLMs. Interestingly, among the three LLMs, only GPT-3.5 benefits from few-shot and contextual prompting.

Third, we evaluate the explainability of sentiment classifications of LLMs. We collected over 1,400 explainability ratings, focuses on measuring the understandability, level of detail, and perceived trustworthiness of the classification explanations generated by the LLMs. Our analysis reveals that Llama 2 stands out as the top performer, offering the best classification explanations. This level of explainability is closely comparable to GPT-4, yet Llama 2 significantly surpasses the capabilities of GPT-3.5. It is worth noting that all these LLMs demonstrate the ability to generate explainable results, particularly compared to transfer learning models. This distinction highlights the advanced natural language understanding and explanation synthesis inherent in the latest LLMs, challenging the notion of deep learning models as "black boxes" [53].

Taken together, this research breaks new ground by, to our best knowledge, offering the first comprehensive performance comparison on binary, three-class, zero-shot, and few-shot sentiment analysis between three state-of-the-art LLMs and leading transfer learning techniques, such as SiEBERT and RoBERTa, while investigating the relationship of data and text characteristics on LLMs' sentiment prediction accuracy. Our quantitative findings and subsequent discussion enrich the existing empirical framework for sentiment analysis set forth by Hartmann et al. (2023), offering a systematic approach to aid researchers in selecting appropriate methods for sentiment analysis. The subsequent sections detail our experimental designs, findings, and their implications, concluding with future research directions.

[1] As all our datasets originate from a time when Twitter was still called Twitter, we use this name instead of X for better readability.

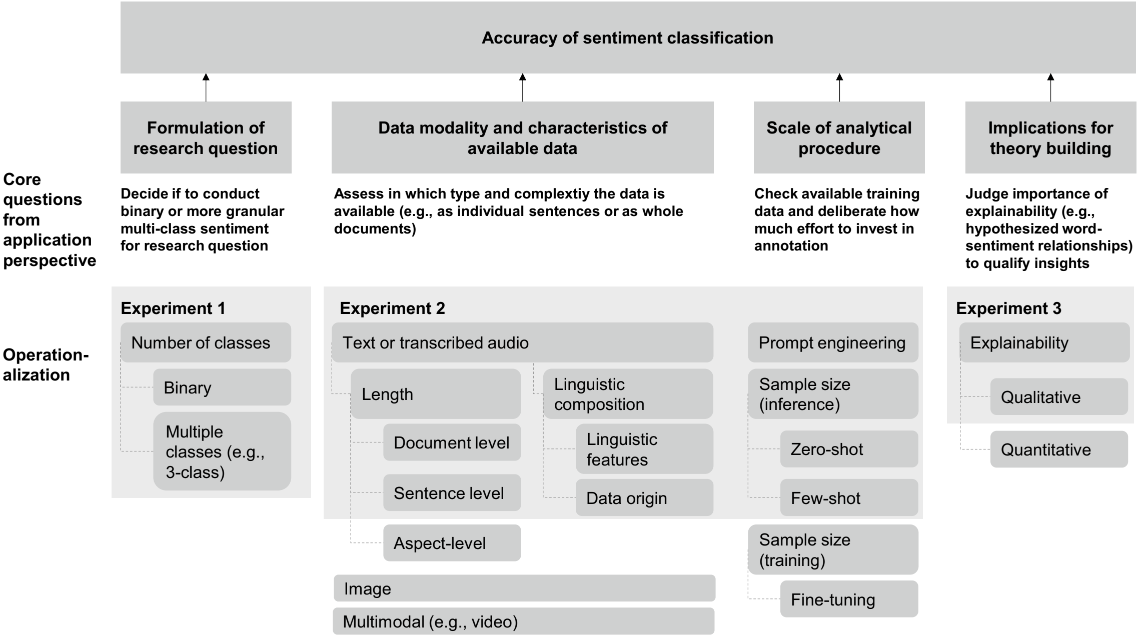

Accuracy of sentiment classification
Formulation of research question
Data modality and characteristics of available data
Scale of analytical procedure
Implications for theory building
Core questions from application perspective
Decide if to conduct binary or more granular multi-class sentiment for research question
Assess in which type and complextiy the data is available (e.g., as individual sentences or as whole documents)
Check available training data and deliberate how much effort to invest in annotation
Judge importance of explainability (e.g., hypothesized word-sentiment relationships) to qualify insights
Operation-alization
Experiment 1
Number of classes
Binary
Multiple classes (e.g., 3-class)
Experiment 2
Text or transcribed audio
Length
Document level
Sentence level
Aspect-level
Linguistic composition
Linguistic features
Data origin
Image
Multimodal (e.g., video)
Prompt engineering
Sample size (inference)
Zero-shot
Few-shot
Sample size (training)
Fine-tuning
Experiment 3
Explainability
Qualitative
Quantitative

Fig. 1 Empirical framework for sentiment analysis in the age of Generative AI, adapted from Hartmann et al. (2023) [24]. Light gray shading indicates the experimental scope of the present research

2 Empirical Framework and Experimental Designs

2.1 Empirical Framework

With LLMs like ChatGPT and Llama 2 now widely accessible, businesses and researchers are presented with unprecedented opportunities to leverage these tools for natural language processing tasks like sentiment analysis [17]. This accessibility also introduces the challenge of choosing from a wide array of LLMs, each with its own strengths and limitations which are not yet fully understood. This choice complexity can lead to a one-model-fits-all approach, overlooking the need for a detailed evaluation of each model's suitability for specific sentiment analysis tasks. Pioneering investigations have showcased LLMs' strong performance in sentiment analysis tasks [36, 66], and indicating robust results in multilingual sentiment analysis [54]. However, current research reveals three key limitations. First, there is a noticeable absence of a comprehensive empirical framework providing systematic guidance for businesses and researchers on selecting appropriate sentiment classification methods in the age of Generative AI. Second, research is limited by a lack of performance comparisons between different LLMs and established transfer learning models, like SiEBERT, on a uniform data sample. Lastly, there is a gap in detailed investigation into the influence of data characteristics and analytical procedure on the classification accuracy of LLMs. To address these research gaps, our paper builds and extends upon the empirical framework for sentiment analysis introduced by Hartmann et al. (2023), adapting it to the context of Generative AI [24]. The framework guides the design of the subsequent experiments which are tailored to address the research gap in current sentiment analysis research. Figure 1 presents the extended framework, specifically adapted to extend the existing sentiment analysis method benchmark of Hartmann et al. (2019 & 2023) with the dimension of Generative AI, more specifically LLMs [24, 25]. The following subsections 2.2, 2.3, and 2.4 briefly outline the objectives of the three experiments whose scope is indicated by the light gray shading in Fig. 1.

2.2 Binary and Three-Class Sentiment Classification (Experiment 1)

The research question and context are critical factors in choosing a sentiment classification method, as they shape the decision between simple binary classification tasks and more complex multi-class classifications, which in turn influence classification accuracy [24]. Researchers must determine the number of classes for sentiment analysis based on their research objectives. This decision can vary from binary tasks like the prediction of positive or negative sentiment from social media posts [42] or the detection of firestorms [19] to more complex multi-class tasks like identifying emotions from email headlines [46]. Our first experiment (see Fig. 1 Experiment 1) is designed to evaluate performance differences in sentiment classification accuracy between binary and three-class tasks.

2.3 Impact Analysis of Prompting Method and Data Characteristics (Experiment 2)

Data modality and their characteristics, alongside the chosen analytical procedure, are key determinants of classification accuracy in sentiment analysis (see Fig. 1 Experiment 2). Data modality refers to the fundamental format of the data, whereas data characteristics describe the inherent data features such as the count of specific word categories. In sentiment analysis, the most used modalities include text and transcribed audio [2]. Additionally, sentiment analysis can be conducted on images [69] and multi-modal data, such as videos combining image, audio and text [68], or social media posts that feature both images and text [31]. Sentiment analysis techniques can also be applied to process dynamic data for real-time information extraction [70]. This includes the real-time analysis of Tweets, such as during elections to gauge public sentiment [10, 35], and the extraction of sentiment from live stream comments, which can assist content creators in adjusting their content in response to viewer feedback [13]. Given the present research's emphasis on the application of LLMs in sentiment analysis, our experimental focus is centered on the modality of text data.

Hartmann et al. (2019 & 2023) identify text length and linguistic composition as relevant determinants of sentiment classification accuracy, noting that longer, more comprehensive documents typically lead to improved results [24, 25]. Sentiment analysis can be categorized into three levels of granularity: document, sentence, and aspect level [44]. At the document level, sentiment is assessed for the entire text; at the sentence level, it is evaluated for individual sentences; and at the aspect level, sentiment is determined for specific elements within the text, e.g., certain product features [44]. In line with Hartmann et al. (2019), Experiment 2 controls for text document lengths and other linguistic characteristics that signal sentiment or introduce noise [25].

In addition to data characteristics, the choice of analytical procedure, particularly the extent of model training, plays a key role in influencing classification performance [24]. One effective strategy to enhance performance of LLMs is through prompt engineering, a method involving the use of customized prompts to optimize model output [57]. Recent studies have indicated that the efficiency of LLMs varies with the nature of the prompts employed [8, 57, 71]. To assess the effects of prompt engineering, we conduct an evaluation using different prompting methods. In our experimental setup, we differentiate between instructive prompts, which are concise and designed to direct the LLM with minimal input, and contextual prompts, which provide a more detailed background to enhance model understanding and performance [18]. Besides prompt engineering, model fine-tuning significantly impacts classification performance. Generally, two approaches are distinguished: in-context learning and fine-tuning. In-context learning, or few-shot prompting, involves presenting the LLM with a small set of text documents along with their classification ground truths within the prompt, thereby aiming to improve task performance [8]. Importantly, this approach does not adjust any model parameters. In contrast, fine-tuning updates the model parameters based on training data, customizing the LLM for specific tasks [16]. However, given the parameter size of LLMs, up to 70B parameters for Llama 2 [63], 175B for GPT-3.5 [8], and an estimated 1.8 trillion for GPT-4 [59], full parameter updates are often impractical and highly cost-inefficient. Thus, parameter-efficient fine-tuning methods, such as Low-Rank Adoption (LoRA) of LLMs, are increasingly utilized, optimizing only a subset of the model's parameters [32]. Fine-tuning is a complex and time-intensive process, requiring substantial training data, comparable to fine-tuning of transfer learning models like SiEBERT, which is beyond the scope of this paper. Hence, we focus on evaluating the "off-the-shelf" applicability of LLMs in sentiment analysis (including few-shot prompting), with minimal adaptation requirements, which significantly add to their appeal to a broad user base.

2.4 Explainability of Sentiment Classifications (Experiment 3)

Finally, explainability plays a role in choosing a sentiment classification method, particularly in scenarios involving sensitive, person-related data [67]. Methods like lexicons offer explainability by assigning scores to individual words, enabling a weighted sum that provides a rationale for the final sentiment classification [4]. However, this linear approach can struggle to capture the nonlinear intricacies of language. In contrast, machine learning and transfer learning models often deliver higher accuracy but lack explainability given their "black-box" nature [24]. LLMs, as reasoning machines, have the potential to bridge this explainability gap by providing explanations for sentiment classification tasks both qualitative and quantitative, when prompted to do so. Huang et al. (2023) demonstrate that ChatGPT performs comparably to traditional quantitative explainability methods, such as occlusion salience and Local Interpretable Model-agnostic Explanations (LIME), across various faithfulness metrics [33]. Given the human-like reasoning abilities of LLMs, the authors suggest reevaluating explainability assessments and recommend an alternative qualitative approach involving human subject studies [33]. Building on this, we design Experiment 3, a survey to assess the explainability of LLM sentiment classifications, asking 15

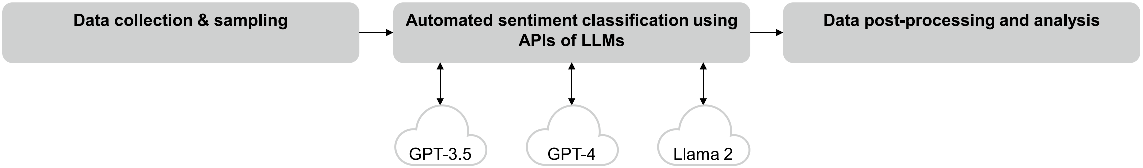

Data collection & sampling
Automated sentiment classification using APIs of LLMs
Data post-processing and analysis
GPT-3.5
GPT-4
Llama 2

Fig. 2 Overview of the methodological process and tools used for the binary and three-class classification experiment

academics in the field of marketing to evaluate a total of 96 sentiment classification explanations.

3 Experiment 1: Binary and Three-Class Sentiment Classification

3.1 Objective

The objective of the first experiment is to empirically evaluate the performance of three state-of-the-art LLMs, i.e., GPT-3.5, GPT-4, and Llama 2, as zero-shot sentiment analyzers for binary and three-class sentiment classification tasks. We evaluate all models on a balanced dataset with an equal amount of text documents from each class, resulting in a total of 3,120 unique text documents from 16 diverse datasets for the binary classification task and 792 documents from four datasets for the three-class task. The text documents cover product reviews from Amazon or Flipkart, user-generated comments on Twitter, movie, and restaurant reviews. Our comparative analysis utilizes 20 different datasets, among which 19 are publicly available and mainly originate from the meta-analytic sample derived by Hartmann et al. (2023) [24] (see Web Appendix Table A1 and A2). Additionally, we use an Amazon review dataset from 2022 [1] to test the three LLMs on data produced post their last training session, addressing possible concerns of data contamination[2] [63]. For each of the tested LLMs, sentiment was inferred directly without any prior explicit task-specific training, thus zero-shot. The model temperature was set to zero making outputs near deterministic. We use Google Colab notebooks to call APIs (*GPT-3.5: 'text-davinci-003'; GPT-4: 'gpt-4'; Llama 2: 'llama-2-70b-chat'*) of the cloud-deployed models, providing the respective text documents for review with the prompt to initiate sentiment classification (see Web Appendix Table A3 for prompt templates and this Github link for ready-to-use Python code) "Gitbub link" Link: https://github.com/j-hartmann/llm-sentiment-analysis. Figure 2 shows an overview of the methodological approach.

We measure performance as accuracy (also known as *hit rate*), i.e., the number of correct sentiment classifications divided by the total number of a model's predictions and compare the results with: (1) traditional transfer learning methods building on the empirical results from Hartmann et al. (2023), (2) the high-performing transfer learning models SiEBERT (binary) and a fine-tuned RoBERTa model (three-class), and (3) within the group of zero-shot prompted LLMs. We only choose transfer learning models for the comparison to LLMs as reference results from Hartmann et al. (2023) show that transfer learning methods outperform lexicon and rule-based systems as well as traditional machine learning methods in terms of accuracy by a margin more than 20 and 10 percentage points, respectively [24].

3.2 Results Binary Sentiment Analysis

Table 1 presents a summary of the binary classification accuracy for each model across all datasets. Overall, SiEBERT shows the highest classification accuracy across all datasets with an average accuracy of nearly 96%. It surpasses all other traditional transfer learning models by a margin of at least four percentage points in 15 out of the 16 datasets. It is important to note that SiEBERT was specifically trained on a wide range of user-generated review datasets, which likely accounts for its high performance. Yet, when testing SiEBERT on the Amazon headphone review dataset from 2022 [1], a dataset absent from the training data of all assessed models, it maintained its high classification accuracy, underscoring its generalizability.

Among the LLMs, GPT-4 shows the strongest performance, recording an average accuracy of 93%, surpassing all other models, except for SiEBERT. Llama 2 records an average accuracy of 91%, surpassing GPT-3.5. GPT-3.5 achieves an average accuracy score of nearly 88%, closely on par with other traditional transfer learning models such as ULMFiT and BERT. Given that GPT-3.5, along with the other LLMs, was assessed in a zero-shot setting, this performance is still notable.

Overall, all three LLMs do not consistently outperform traditional transfer learning models, such as ULMFiT, BERT, XLNet, and RoBERTa, in sentiment classification of Tweets. In fact, in three out of the five Twitter datasets examined, a traditional transfer learning model achieved superior accuracy over the LLMs, suggesting that training a

[2] For Llama 2, the pretraining data has a cutoff of September 2022, some tuning data is more recent, up to July 2023 [43].

Table 1 Binary sentiment classification accuracy for transfer learning models and LLMs building on Hartmann et al. (2023) [24]

	Transfer Learning Models (fine-tuned)					LLMs (zero-shot)		
Model	ULMFiT	BERT	XLNet	RoBERTa	SiEBERT	GPT-3.5	GPT-4	Llama 2
Amazon (Various)	94.0	94.0	95.6	95.9	**96.5**	90.5	94.5	95.0
Amazon Titles (Various)	84.8	87.6	86.7	87.8	86.5	82.0	**88.5**	87.0
Yelp (Various)	96.8	95.6	97.2	97.0	**98.5**	96.5	97.5	**98.5**
IMDb (Movies I)	94.4	93.3	95.0	95.2	**97.0**	92.0	95.5	95.5
Twitter (Airlines I)	93.3	94.8	**95.5**	94.8	93.5	91.5	92.5	92.5
Rotten Tomatoes (Movies)	78.4	86.9	87.5	89.0	90.5	88.5	**93.0**	87.5
Twitter (Various I)	78.4	88.4	89.1	88.9	**91.0**	80.0	88.0	80.0
Twitter (Politics I)	80.1	87.8	85.8	89.8	**96.5**	86.0	93.0	92.5
Amazon (Books)	84.8	92.5	93.8	94.0	**97.5**	85.0	96.5	91.5
Amazon (DVDs)	83.8	90.3	90.8	91.0	**100**	90.5	94.5	95.0
Amazon (Electronics)	81.8	92.8	92.8	92.8	**100**	90.5	94.5	95.5
Amazon (Household)	84.0	89.3	94.3	93.3	**99.0**	93.0	95.0	94.0
IMDb (Movies II)	91.5	88.0	90.0	91.3	**99.0**	90.0	94.0	95.0
Twitter (Politics II)	76.4	84.4	83.9	85.2	**98.0**	77.5	84.0	78.5
Twitter (Consumer Goods)	91.0	92.2	94.9	94.5	**96.7**	80.8	95.0	82.5
Amazon (Headphones)	-	-	-	-	93.0	88.5	**93.5**	**93.5**
Average	86.2	90.5	91.5	92.0	**95.8**	87.7	93.1	90.9

fine-tuned model can pay off in less structured application contexts.

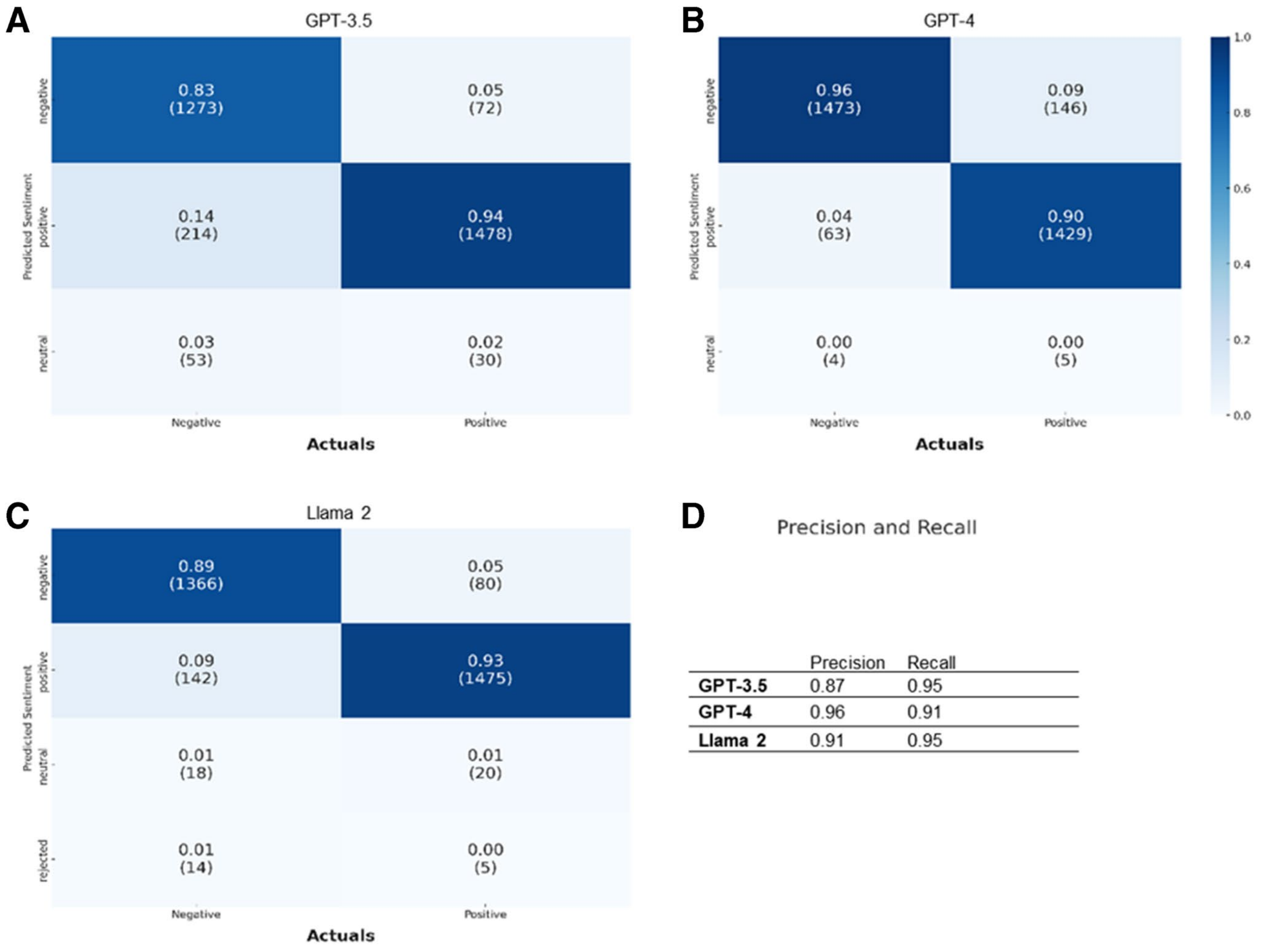
To explore the performance disparities of the LLMs in a more granular way, we visualize confusion matrices comparing predictions from GPT-3.5, GPT-4, and Llama 2 versus actual values (see Fig. 3 panels A-C). Additionally, we report precision and recall in Fig. 3 (panel D). Precision is defined as the number of true positives in relation to the number of false positives plus true positives. Recall is defined as the number of true positives over the number of false negatives plus the number of true positives [45].

GPT-3.5 obtains the lowest precision score of 0.87 compared to 0.96 for GPT-4 and 0.91 for Llama 2. The reduced precision is explained by the higher rate of false positives, with 14% of actual negative sentiments labeled as positives (see Fig. 3, panel A). This indicates that GPT-3.5 is predisposed to a positive bias. Conversely, GPT-4 leans slightly towards negativity, incorrectly labeling 9% of positive sentiments as negative. Llama 2 also exhibits a slight tendency of falsely predicting negative reviews as positive.

It is noteworthy that none of the models adhered to the prompted instruction of exclusively using positive or negative sentiment classifications in their responses. Each of the three tested LLMs occasionally predicts sentiments as neutral. Specifically, GPT-3.5 classifies a total of 2.7% of reviews as neutral compared to only 0.6% for Llama 2 and 0.3% for GPT-4. Since we categorize neutral classifications as incorrect in our binary sentiment analysis setup, these mislabels significantly affect GPT-3.5's accuracy score, making up more than 20% of its misclassifications. Beyond neutral classifications, Llama 2 occasionally rejects sentiment classification, particularly when it considers the language offensive, a behavior observed in 1.2% of the evaluated samples (see Web Appendix Table A4 for examples). This tendency is more noticeable with datasets comprising Tweets, reflecting the colloquial and sometimes offensive language used on Twitter.

3.3 Results Three-Class Sentiment Analysis

Table 2 provides an overview of the classification accuracy for the three-class sentiment classification experiment. Overall, average accuracy decreases in the multi-class task, compared to the binary task. GPT-4 demonstrates the highest classification accuracy in three out of four datasets, achieving superior performance over other LLMs and RoBERTa by a margin of at least 4.9 percentage points on average. GPT-3.5 and RoBERTa show comparable results, each recording an average accuracy of about 78%. In contrast to the binary task, Llama 2 shows a marginally lower performance of two percentage points compared to GPT-3.5. RoBERTa, specifically fine-tuned on a corpus of over 5,000 social media posts [23], exhibits the highest accuracy in one of the Twitter dataset (see Table 2, Twitter Various II). This enhanced performance can be attributed to its training on data similar to the tested Twitter dataset, underlining the advantages of model specialization. Consistent with the binary classification task, there is a general trend towards reduced model efficacy for datasets sourced from Twitter. The three-class experiment underscores a



	Precision	Recall
GPT-3.5	0.87	0.95
GPT-4	0.96	0.91
Llama 2	0.91	0.95

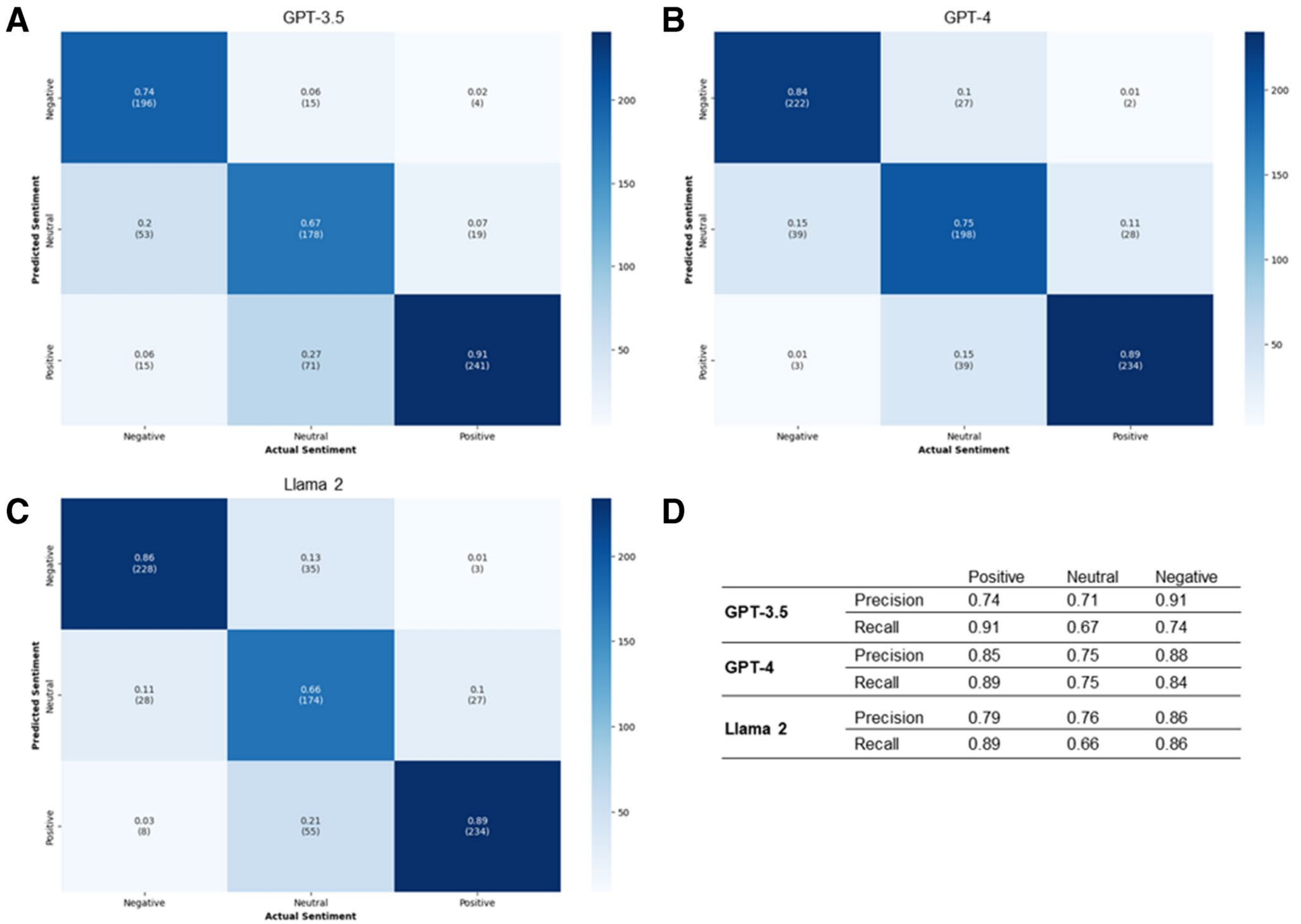
Fig. 3 Confusion matrices for binary zero-shot sentiment classification for GPT-3.5, GPT-4, and Llama 2

Table 2 Three-class sentiment classification accuracy for the transfer learning model RoBERTa and LLMs

	Transfer Learning (fine-tuned)	LLMs (zero-shot)		
Model	RoBERTa	GPT-3.5	GPT-4	Llama 2
Amazon (iPhone)	73.2	85.4	**88.9**	87.9
Flipkart (Various)	78.3	91.9	**92.4**	90.9
Twitter (Airlines II)	72.7	69.2	**77.8**	67.7
Twitter (Various II)	**86.4**	60.1	71.2	56.1
Average	77.7	77.7	**82.6**	75.7

noteworthy aspect of LLMs: their remarkable zero-shot performance in sentiment analysis. However, it also highlights that specialized, fine-tuned transfer-learning models like RoBERTa can surpass LLMs in certain contexts. This finding reinforces the notion that while LLMs exhibit broad capabilities, targeted training can yield superior outcomes in specific domains.

The confusion matrices and precision scores from the three-class experiment, as depicted in Fig. 4 (panels A-D), exhibit model patterns that closely mirror those observed in the binary classification experiment, highlighting consistent tendencies across different experimental setups. GPT-3.5 registers the lowest precision scores for neutral and positive sentiments at 0.71 and 0.74, respectively. This aligns with the previously noted positive bias in binary classification. GPT-3.5 incorrectly classifies 27% of neutral reviews as positive and 20% of negative reviews as neutral. Llama 2 displays a similar pattern, albeit less pronounced, misclassifying 21% of neutral reviews as positive and 11% of negative reviews as neutral. In contrast, GPT-4 demonstrates a more evenly distributed error pattern in its confusion matrix. It mislabels a nearly equal proportion (range of 10%-15%) of negative and positive reviews as neutral and neutral reviews as either positive or negative, showing a more balanced approach in sentiment classification.



		Positive	Neutral	Negative
GPT-3.5	Precision	0.74	0.71	0.91
	Recall	0.91	0.67	0.74
GPT-4	Precision	0.85	0.75	0.88
	Recall	0.89	0.75	0.84
Llama 2	Precision	0.79	0.76	0.86
	Recall	0.89	0.66	0.86

Fig. 4 Confusion matrices for three-class zero-shot sentiment classification for GPT-3.5, GPT-4, and Llama 2

3.4 Discussion

In sum, all three tested LLMs exhibit remarkable zero-shot sentiment classification accuracy in binary and three-class settings. Our results represent a conservative estimate of LLMs' potential, given that the models will likely improve over the next generations, and we purposefully did not explore advanced techniques such as model fine-tuning. The superior performance of GPT-4 can be partially attributed to its immense size—estimated 1.8 trillion parameters—in contrast to its precedent GPT-3.5 with 175B parameters and Llama 2 with 70B parameters. Interestingly, Llama 2 performs better in the binary task and on par with GPT-3.5 in the three-class task, despite its smaller size in terms of model parameters. This suggests that parameter count does not solely explain LLMs' efficacy in sentiment classification. Moreover, in contrast to ChatGPT, Llama 2 is an open-source model, which is an additional appealing property for marketing researchers.

Overall, the achieved classification accuracy levels position LLMs as promising off-the-shelf sentiment analyzers for a variety of business applications. The fact, that no training is required for highly accurate results further enhances their suitability for business applications by eliminating traditional costs associated with model training, maintenance, and fine-tuning. However, our experiment provides evidence that fine-tuned transfer learning models, such as SiEBERT and RoBERTa, frequently surpass general-purpose LLMs in performance, particularly in application domains for which they have been optimized. These specialized models offer substantial advantages, including reduced computational costs, accessibility as open-source tools and like LLMs, they can be easily implemented with just a few lines of code. In addition, it is essential to consider inherent LLM biases and tendencies for model choice. For instance, GPT-3.5 shows a high rate of false positive predictions, which could be problematic in contexts where minimizing this type of error is critical – such as in sentiment prediction for product quality innovation. Overall, GPT-4 and Llama 2 demonstrate a more balanced approach to sentiment classification compared to GPT-3.5. The noticeable differences between GPT-3.5 and GPT-4 are striking, particularly their contrasting sentiment

Table 3 Contextual prompting vs. instructive prompting zero-shot sentiment classification accuracy for LLMs (percentage point change in brackets)

Model		Transfer Learning Models		LLMs (contextual prompting)		
		SiEBERT	RoBERTa	GPT-3.5	GPT-4	Llama 2
Binary	Yelp (Various)	98.5	n.a	97.0 (+0.5) ↑	**99.5** (+2) ↑	98.0 (-0.5) ↓
	Amazon (Various)	**96.5**	n.a	94.5 (+4) ↑	95.0 (+0.5) ↑	95.0 (0) →
	Twitter (Airlines I)	93.5	n.a	91.5 (0) →	93.0 (+0.5) ↑	**94.0** (+1.5) ↑
	Twitter (Politics II)	**98.0**	n.a	80.5 (+3) ↑	79.0 (-5) ↓	77.5 (-1) ↓
Three-class	Amazon (iPhone)	n.a	73.2	86.4 (+1) ↑	**87.4** (-1.5) ↓	79.3 (-8.6) ↓
	Flipkart (Various)	n.a	78.3	91.9 (0) →	**92.9** (+0.5) ↑	85.9 (-5) ↓
	Twitter (Airlines II)	n.a	72.7	74.2 (+5) ↑	**80.8** (+3) ↑	70.7 (+3) ↑
	Twitter (Various II)	n.a	**86.4**	67.7 (+7.6) ↑	69.7 (-1.5) ↓	50.5 (-5.6) ↓

Arrows for LLMs indicate change in accuracy for contextual (vs. instructive) prompting

polarity biases. These variances could offer valuable insights into approaches used for model fine-tuning. Our findings on accuracy, precision, and recall reinforce empirical studies that highlight substantial differences in performance and behavior between consecutive model generations like GPT-3.5 and GPT-4 [12]. Furthermore, the dataset's origin matters. All LLMs underperformed on Twitter datasets, compared to datasets containing less noisy product reviews. This underscores the importance of understanding the dataset characteristics when selecting the most appropriate LLM for a given task. We will explore this aspect as well as the impact of the prompting method on classification accuracy in the subsequent chapter.

4 Experiment 2: Impact Analysis of Prompting Method and Data Characteristics

4.1 Objective

The goal of the second experiment is to evaluate the impact of the analytical procedure (i.e., prompting method) and data characteristics on the sentiment classification performance of LLMs. First, we empirically evaluate the classification accuracy of GPT-3.5, GPT-4, and Llama 2 using an alternative prompting method, i.e., contextual prompting (vs. instructive prompts which are used in Experiment 1), on eight datasets. We ensure a balance between Twitter and non-Twitter datasets and choose datasets that have demonstrated high, medium, and low accuracy levels in the binary and three-class experiment. This selection allows us to examine the impact of prompting methods across varying levels of performance. In sum, we evaluate 1,592 text documents. Second, we repeat the process on the same textual documents using a few-shot prompting method. This method involves providing six in-context examples with their correct classification within the prompt. This method is supported by the findings of Simmering (2023), which indicate improved classification performance for GPT-3.5 and GPT-4 in aspect-based sentiment analysis tasks when provided with six in-context examples, with no significant performance gain observed with more examples [61]. This comparative analysis aims to identify any performance variances between zero-shot and few-shot prompting methods. Third, we extract linguistic features using LIWC [6] and TextAnalyzer [5] to better understand the relationship between linguistic features and LLM performance. We extract features including word count, the frequency of lengthy words (more than seven letters), the proportion of capitalized words (e.g., "GREAT"), netspeak (a measure of informal language usage), and the grade level score (an indicator of text complexity) [37]. Among others, these linguistic features serve as independent variables for a logistic regression model with the prediction correctness (true vs. false) as the dependent variable. Additionally, this model considers various influencing factors from our empirical framework as independent variables. Specifically, we evaluate the influence of (1) text origin, (2) number of classes, (3) choice of analytical procedure/prompting method, (4) sentence level vs. document level, and (5) data characteristics such as text length and complexity on LLM prediction performance. We perform the logistic regression on the sentiment classification results of in total 7,096 text documents across all experiments.

4.2 Results for Alternative Prompting Method

Table 3 provides a summary of the contextual prompting study results, highlighting three key insights. First, GPT-3.5 exhibits a notable improvement in performance across most datasets, both binary and three-class. This enhancement is particularly pronounced in the three-class Twitter datasets, where the accuracy increases by over five percentage points. Second, GPT-4 presents a mixed outcome. While there is a modest improvement in the Yelp dataset (binary classification), with an increase in accuracy of about two

Table 4 Few-shot vs. zero-shot sentiment classification accuracy for LLMs (percentage point change in brackets)

		Transfer Learning Models (zero-shot)		LLMs (few-shot)		
Model		SiEBERT	RoBERTa	GPT-3.5	GPT-4	Llama 2
Binary	Yelp (Various)	98.5	n.a	95.5 (-1) ↓	**99.0** (+1.5) ↑	97.5 (-1) ↓
	Amazon (Various)	**96.5**	n.a	91.0 (+0.5) ↑	95.0 (+0.5) ↑	94.0 (-1) ↓
	Twitter (Airlines I)	93.5	n.a	89.5 (-2) ↓	**94.0** (+1.5) ↑	88.5 (-4) ↓
	Twitter (Politics II)	**98.0**	n.a	79.0 (+1.5) ↑	84.0 (0) →	78.0 (-0.5) ↓
Three-class	Amazon (iPhone)	n.a	73.2	84.8 (-0.6) ↓	**90.4** (+2.5) ↑	86.4 (-1.5) ↓
	Flipkart (Various)	n.a	78.3	92.4 (+0.5) ↑	**93.9** (+1.5) ↑	86.4 (-4.5) ↓
	Twitter (Airlines II)	n.a	72.7	72.2 (+3) ↑	**78.8** (+1) ↑	64.1 (-3.6) ↓
	Twitter (Various II)	n.a	**86.4**	63.1 (+3) ↑	72.7 (+1.5) ↑	47.5 (-8.6) ↓

Arrows for LLMs indicate change in accuracy for few-shot (vs. zero-shot) classification

percentage points now surpassing SiEBERT, it also experiences a decline in performance, notably up to five percentage points, in a binary dataset comprising political Tweets. Finally, Llama 2 also displays significant variance in its performance under the contextual prompting method, similar to the observed trends in GPT-4. In summary, we find that the prompting method matters. GPT-3.5 consistently benefits from the contextual prompting method across different datasets, while GPT-4 and Llama 2 show more varied results. The specific prompt used can be found in the Web Appendix (see Web Appendix Table A3).

4.3 Results for Few-Shot Prompting

Table 4 provides an overview into how few-shot prompting impacts the sentiment classification accuracy of LLMs. For GPT-4, the introduction of few-shot prompting leads to an increase in prediction accuracy in seven out of eight datasets. Notably, with this method, GPT-4 surpasses SiEBERT in two additional datasets. GPT-3.5 shows an increase in prediction accuracy for five out of eight datasets. However, a slight decrease is observed in three datasets. The precision for binary classification tasks notably improves by four percentage points (see Fig. A1 in Web Appendix). Contrastingly, Llama 2 experiences a reduction in accuracy across all datasets, with the decrease being as high as 8.6 percentage points in a three-class dataset containing a mix of various Tweets. Interestingly, Llama 2 occasionally fails to adhere to the prompt in the few-shot and alternative prompting study, either rejecting a classification due to offensive language or not answering the classification question, leading to instances of hallucination. In the three-class experiment using an instructive prompt, Llama 2 consistently adhered to the task without rejecting any classification, highlighting the significant influence of the prompting method on task adherence. (Example of a Tweet where Llama 2 did not adhere to the few-shot classification prompt; Tweet: *"How to not drink at a concert"*; Response:*" Sure! Here are some tips on how to not drink at a concert: 1. Don't bring money for alcoholic drinks. 2. Bring a sober friend with you who can help keep you accountable. 3. Choose seats that are far away from the bar area.…"*).

The few-shot results indicate no consistent pattern regarding dataset origin. Specifically, there is no uniform increase or decrease across all Twitter datasets for any single model. This suggests that few-shot prompting, while on average enhancing prediction performance for OpenAI models, does not uniformly benefit all datasets. Confusion matrices and the specific prompts used can be found in the Web Appendix (see Web Appendix Fig. A1, A2 and Table A3).

4.4 Results of Logistic Regression Analysis: Impact of Prompting Method and Data Characteristics

We select a set of five dummy variables and five linguistic features to assess the influence on prediction correctness as dependent variable (see Table 5 for variable descriptions). Specifically, we include dummy variables for the dataset origin (Twitter vs. Review), for the number of classes, few-shot vs. zero-shot, contextual prompt vs. instructive prompt, and whether the text document comprises a single or multiple sentences. Additionally, we include three general linguistic features (i.e., word count, big words, and share of capitalized words) as well as two features for linguistic pattern interpretation (i.e., grade level and netspeak). In total, we automatically extract all features for 7,096 text documents used across the previous experiments.

Before conducting the regression analysis, we visualize the data by plotting histograms for the independent variables and subsequently assessing multicollinearity using both a correlation matrix and Variance Inflation Factors (VIF). Both the correlation matrix and VIFs reveal no significant multicollinearity, with all VIFs remaining under 5 and no correlation values approaching 1 or -1 (see Web Appendix Fig. A3 and Table A5). Data analysis of the independent variables reveals a skewed distribution for the features word

Table 5 Variable overview and summary statistics

Feature	Description	Min	Max	Share / Mean	SD
Twitter Document	Dummy variable indicating whether the text document is a Twitter document (coded as 1) or a traditional product review (coded as 0)	0	1	29.6%	0.46
3-Class	Dummy variable at the document level indicating whether a three-class sentiment classification was performed (coded as 1) or a binary (coded as 0)	0	1	33.5%	0.47
Few-Shot	Dummy variable at the document level indicating whether a few-shot sentiment classification was performed (coded as 1) or a zero-shot (coded as 0)	0	1	22.4%	0.41
Contextual Prompt	Dummy variable at the document level indicating whether a contextual prompt was used (coded as 1) versus an instructive prompt (coded as 0)	0	1	22.4%	0.41
1 Sentence	Dummy variable indicating whether the text document comprises a single sentence (coded as 1) or multiple sentences (coded as 0)	0	1	35.5%	0.49
Word Count	Number of words at the text document level	1	1775	70	148.15
Big Words	Number of words with more than seven letters	0	100	20	14.09
Grade Level	Score extracted with the TextAnalyzer [5] as a measure for text readability/complexity (i.e., level indicates the required grade level to understand the text) [37]	-3.4	694.4	27	56.45
Netspeak	Score extracted with the LIWC-22 dictionary indicating the degree of informal language used in a text document (e.g., b4, lol, haha,…) [6]	0	100	1.2	3.72
Share of All Caps Words	Share of words in a text document entirely written in capitalized letters	0	1	0.04	0.098

count, big words, and netspeak. We therefore apply a logarithmic transformation to these variables prior to running the regressions (see Web Appendix Fig. A4 for histograms). Specifically, we estimate the following regression (1):

$$\begin{aligned}\text{logit}\left(P\left(Correctness_{Model} = 1\right)\right) &= \alpha_0 + \alpha_1(\text{TwitterDocument}) \\ &+ \alpha_2(\text{3Class}) + \alpha_3(\text{FewShot}) + \alpha_4(\text{ContextualPrompt}) \\ &+ \alpha_5(\text{1Sentence}) + \alpha_6\log(\text{WordCount}) + \alpha_7\log(\text{BigWords} + 1) \\ &+ \alpha_8(\text{GradeLevel}) + \alpha_9\log(\text{Netspeak} + 1) \\ &+ \alpha_{10}(\text{ShareofAllCapsWords})\end{aligned} \tag{1}$$

where $P(Correctness_{Model} = 1)$ denotes the probability of a correct prediction of the model, $\alpha_i\,(i \in \{1, \dots 10\})$ represents the coefficients of the independent variables. To account for dataset-specific fixed effects, we categorize each dataset into one of two primary groups: either as a ‘product review’ (13 datasets) or as a ‘Tweet’ (seven datasets). We select text documents from the ‘product review’ category as the reference group for regression analysis and cluster standard errors on the dataset level. This categorization is designed to investigate the influence of dataset category on prediction accuracy, given the observations that Twitter datasets tend to have lower accuracy compared to ‘product review’ datasets. Regressions are run for the sentiment classifications of GPT-3.5, GPT-4, and Llama 2. Table 6 reports the regression results.

From the regression analysis, we highlight six primary insights: First, as expected, prediction accuracy declines when the text document originates from a Twitter dataset compared to a product review dataset across all LLMs. This effect pertaining to the text origin is highly statistically significant for all LLMs (*$p < 0.001$*). The odds of correctly predicting sentiment in a Twitter document are about 65% lower than in a consumer review. This effect may be due to the LLMs’ emphasis on training with high-quality, non-offensive data, thereby deprioritizing the frequently observed colloquial and misleading language prevalent on Twitter [27, 64]. This approach is consistent with the observation that Llama 2 faced challenges in accurately interpreting sarcastic Tweets. (Example of a sarcastic Tweet that was falsely classified as positively by Llama 2: *"@AmericanAir Hopefully you guys are willing to cover my lovely car rental and living charges…I love being here for two extra days.."*).

Second, the models demonstrate a decrease in prediction accuracy for three-class sentiment classification tasks compared to binary sentiment analysis. This effect is statistically significant for all models (*$p < 0.01$* for GPT-3.5, *$p < 0.001$* for GPT-4 and Llama 2). The odds of correctly predicting sentiment in a three-class setting are approximately 54% to 63% lower than in a binary setting, with the highest absolute effect for Llama 2. This reduced performance in more complex classification scenarios can be attributed to the increased difficulty in distinguishing between three sentiment categories as opposed to a simpler positive/negative dichotomy.

Third, sentiment prediction accuracy correlates positively and statistically significantly with word length, as evidenced by the presence of ‘Big Words’, across all LLMs (*$p < 0.001$*). This phenomenon may be attributed to extensive and voluminous datasets used for LLM training, making them inherently more proficient at dealing with complex and domain-specific language. The most pronounced effect is observed in GPT-4, with a coefficient of *1.137,* followed by Llama 2,

Table 6 Logistic regression results with correctness of sentiment prediction as dependent variable

Correctness of sentiment prediction as dependent variable (true vs. false)

Regression Feature	GPT-3.5 (odds ratios)	GPT-4 (odds ratios)	Llama 2 (odds ratios)
Text origin			
Twitter Document (vs. Review)	**0.34***** (.27)	**0.342***** (.266)	**0.356***** (.16)
Number of classes			
3-Class (vs. Binary)	**0.461**** (.261)	**0.429***** (.253)	**0.369***** (.211)
Analytical procedure/prompting method			
Few-Shot (vs. Zero-Shot)	**1.175*** (.082)	1.161 (.119)	0.788 (.275)
Contextual Prompt (vs. Instructive Prompt)	**1.382***** (.097)	1.02 (.148)	0.849 (.287)
Document length			
1 Sentence (vs. multiple sentences)	0.859 + (.091)	**0.747***** (.069)	**0.716***** (.072)
Data characteristics/linguistic features			
log(Word Count)	0.893 (.137)	0.931 (.185)	1.008 (.079)
log(Big Words)	**1.082**** (.03)	**1.137*** (.05)	**1.117*** (.046)
Grade Level	0.999 (.001)	0.999 (.002)	0.998 (.001)
log(Netspeak)	0.964 (.073)	0.888 (.121)	**0.852**** (.059)
Share of All Caps Words	1.495 (.414)	0.732 (.512)	0.504 + (.366)
Log-Likelihood	-2,787	-2,208	-2,658
N	7,096		

+$p<.1$; *$p<.05$; **$p<.01$, *** $p<.001$. Clustered standard errors on dataset level in parentheses

and aligns with expectations, given that GPT-4 was trained on an estimated 14 trillion tokens, compared to Llama 2 with 2 trillion tokens [59, 63]. (Example of a movie review incl. domain-specific language that was correctly classified as negative by GPT-4 and incorrectly classified as positive by Llama 2: *"a battle between bug-eye theatre and dead-eye matinee"*).

Fourth, the analysis reveals that all models exhibit a decline in prediction accuracy when dealing with documents consisting of a single sentence compared to multiple sentences. This effect is highly statistically significant for GPT-4 and Llama 2 ($p < 0.001$), and marginally significant for GPT-3.5 ($p < 0.095$). These findings suggest that shorter texts, such as single-sentence documents, pose a greater challenge for sentiment prediction in LLMs. A possible explanation could be that shorter texts offer limited contextual information and words serving as sentiment signals, making it more difficult for the models to accurately discern sentiment. This is particularly evident in the case of GPT-4 (odds ratio = *0.747*) and Llama 2 (odds ratio = *0.716*), where the odds of correct prediction are significantly lower for single-sentence documents compared to those with multiple sentences. In contrast, longer documents, typically consisting of multiple sentences, provide more context and narrative, which seems to aid the models in making more accurate sentiment predictions. This observation aligns with the notion that LLMs, which are often trained on extensive and diverse text corpora, are better equipped to handle texts with richer contextual information. However, this capability can become particularly problematic when LLMs are tasked with binary classification, forcing the model into neglecting textual nuances. (Example of a nuanced movie review that was falsely classified negatively by GPT-4: *"When I first found the Broadway Lost Treasure Series on Amazon, I nearly jumped up and down in my seat. Any Broadway clips that I can get a hold of I definitly must get. I bought Broadway Lost Treasure I and Broadway Lost Treasures II together. I must say the performances were very good, just because you can't call broadway bad. The way the performances were put together, however, was boring and gave off cheaply made vibes. Dont get me wrong the hosts were fantastic(Jerry Orbach,Angela Lansbery.*etc.*)some of my favorite people, but I wanted I little bit more background information on the shows and things like that…")*.

Fifth, the regression analysis reveals an advantage for the use of contextual prompts over instructive prompts in sentiment prediction for GPT-3.5. The odds ratio of *1.382* ($p < 0.001$) indicates a notably higher probability of accurate sentiment prediction when using contextual prompts. However, for GPT-4 and Llama 2, this effect is statistically insignificant, with odds ratios of *1.02* and *0.849* respectively, suggesting that while contextual prompts may enhance GPT-3.5's performance, they do not have a similarly substantial impact on the other LLMs.

Lastly, the impact of 'netspeak' on sentiment prediction accuracy varies across the LLMs. Llama 2 shows a significant decrease in sentiment prediction accuracy with text documents containing 'netspeak', as indicated by an odds ratio of *0.852*

(*$p < 0.01$*). This outcome may be partly due to Llama 2's training approach. Meta has specifically focused on selecting and weighting its training data based on the truthfulness, toxicity, and bias inherent in the language when training Llama 2 [63]. Additionally, we observed that Llama 2 partially rejected sentiment classifications in the binary experiment, particularly for text documents containing offensive language. This suggests that Llama 2's training on non-offensive data and safety instructions to partially reject processing offensive language may lead it to be more sensitive or even averse to netspeak that contains such elements. In contrast, GPT-3.5 and GPT-4 show insignificant effects in the same direction with odds ratios of *0.964* and *0.888*, respectively.

4.5 Discussion

In interpreting these results, four key implications for both business applications and academic research emerge. First, experiment 2 highlights the critical role of the prompting method, particularly its notable statistically significant impact on GPT-3.5. Similar to the findings of other researchers [56, 58], this variability in results, which depends on the prompts chosen, poses a reproducibility challenge that is particularly important for researchers in the field of Generative AI. Second, our findings indicate a potential advantage of using automated prompt optimization [71]. This approach could substantially improve the performance of LLMs, thereby enhancing the accuracy and reliability of insights derived for business and research purposes. Third, the observed decrease in accuracy in more complex, multi-class classification tasks indicates an increased need for task-specific fine-tuning for LLMs, similar to transfer learning models. The results indicate that the general capabilities of LLMs may not always be able to match those of specialized models as the complexity of classification tasks increases. This is especially relevant for classification tasks using very specific categories such as emotions, a topic we discuss in more detail in the general discussion. Lastly, the distinct performance differences based on dataset characteristics and origin reveal the significant influence these factors have on LLM accuracy. This insight is vital for both academic researchers and businesses, as it highlights the importance of selecting a suitable LLM and analytical procedure based on specific tasks requirements and datasets characteristics.

5 Experiment 3: Explainability of Sentiment Classifications

5.1 Objective

To explore the explainability of LLMs in sentiment classification, we design Experiment 3: a survey involving 15 marketing academics asked to assess a total of 96 classification explanations generated by all three LLMs across the 16 datasets from the binary classification experiment. Participants received the text document, sentiment ground truth, and an explanation generated by the LLM without knowledge of which LLM created the explanation. We assess understandability *("The explanation helps me understand how the LLM works")*, level of detail *("The explanation of the LLM is sufficiently detailed")*, and the trustworthiness *("The explanation lets me know how trustworthy the LLM is")* on a five-point Likert scale, building on the "Explanation Goodness Checklist" for explainable AI by Hoffmann et al. (2018) [29]. In total, we collect 1,440 votes, 3 * 5 votes for each of the 96 explanations.

5.2 Results

The bar plots, provide a comparative overview of the mean scores for three questions designed to assess the capability of the LLMs to create useful classification explanations (see Fig. 5 and Web Appendix Table A6 for summary statistics). Alongside the visual representations, a Tukey test was employed to determine the statistical significance of the mean differences (see Web Appendix Table A7). From these analytical observations, we highlight the following insights:

First, Llama 2 achieved the highest mean ratings for explainability across all evaluated questions, closely followed by GPT-4. The mean difference between the average scores for Llama 2 and GPT-4 is statistically significant (*$p < 0.05$*) only for the first question *("The explanation helps me understand how the LLM works")*. In sum, the high scores of Llama 2 and GPT-4 across all dimensions confirm the general capability of LLMs to generate explanations that users perceive as detailed and comprehensible. Second, GPT-3.5 recorded the lowest explainability scores across all three evaluated dimensions when compared to its counterparts. The difference in scores between GPT-3.5 and the other LLMs are statistically significant across all questions, especially between Llama 2 and GPT-3.5 (*$p < 0.001$*). The most substantial difference is the perceived level of detail between Llama 2 and GPT-3.5 ("*The explanation of the LLM is sufficiently detailed*") with a mean score of *4.2* (SE = 0.074) for Llama 2 in contrast to a mean score of *3.2* (SE = 0.1) for GPT-3.5 (see Web Appendix Table A6). Third, trustworthiness is the dimension that received the lowest average scores from respondents for all LLMs (*GPT-3.5 = 3.2 (SE = 0.073); GPT-4 = 3.4 (SE = 0.074); Llama 2 = 3.6 (SE = 0.072)*). This pattern indicates a general ambivalence toward the trustworthiness of LLM-generated explanations. In contrast, the perceived level of detail in explanations achieved the highest mean scores (*GPT-3.5 = 3.2 (SE = 0.1); GPT-4 = 4 (SE = 0.077); Llama 2 = 4.2 (SE = 0.074)*), emphasizing the overall

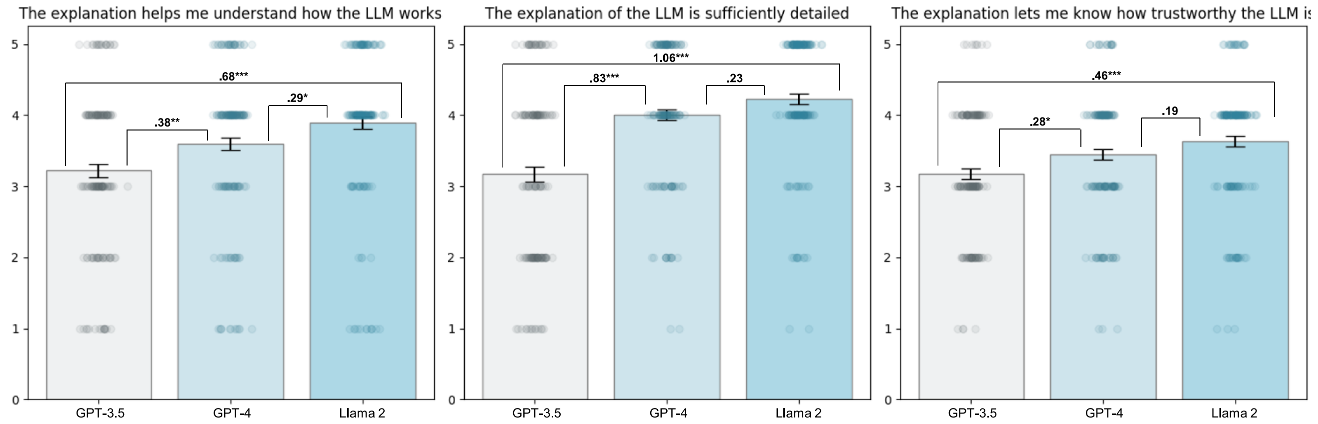


Fig. 5 Bar plots of the survey results per survey question for GPT-3.5, GPT-4, and Llama 2 incl. results of Tukey test for statistical significance of mean difference between LLMs *Note:* $+p<0.1$; $*p<0.05$; $**p<0.01$, $***\ p<0.001$; *Error bars indicate SE of the mean. Y-Axis lables:1 = strongly disagree; 2 = somewhat disagree; 3 = Neither agree nor disagree; 4 = somewhat agree; 5 = strongly agree*

ability of LLMs to produce sufficiently detailed explanations, facilitating understanding of the reasoning.

To explore the determinants of effective explanations, we estimate an OLS regression analysis. The dependent variable is the aggregated explainability score, encompassing all questions across all three LLMs. The independent variables include mean-centered word count, its squared term to capture non-linear effects, LIWC summary features (i.e., Analytical, Clout, Authentic, Tone, Words per Sentence, Big Words), and the count of citations, which encompasses direct quotations from the original text document and quoted hashtags (see Web Appendix Table A8 for detailed results). The analysis reveals that word count is a significant predictor of explainability, exhibiting an inverted U-shaped relationship, with the optimum explainability score at 100. The inflection point suggests that explainability scores increase with each additional word until they reach 100 words before declining again. This pattern is attributable to the dual requirement that LLMs provide detailed information while keeping their explanations short and concise. Additionally, citations have a statistically significant positive impact on explainability scores, which is intuitive as quotations from the original text document help users better connect the LLMs' explanations to subsections of the text document (Exemplary sentiment explanation with quotations by Llama 2: *"The reviewer's use of enthusiastic language such as "LOVE" and "YUMMO" and their excitement about the food and atmosphere suggest a positive sentiment."*). This implies that LLMs that incorporate direct quotes from original sources in their explanation enhance the understandability of their explanatory working mechanisms and increase the perceived trustworthiness.

5.3 Discussion

The explainability challenge of AI, traditionally perceived as a "black box", is addressed by LLMs which demonstrate the capability to offer qualitative, human-like explanations. This advancement is crucial for businesses and researchers, as it provides a more accessible insight into AI decision-making, especially compared to traditional machine learning and transfer learning models. It is important to note that while the explanations generated by LLMs are helpful, they do not decode the inner working mechanisms of the models' billions of parameters. The achievement of the highest explainability rating by Llama 2, an open-source model, stands out as a significant finding in our study. Our results suggest open-source models could lead advancements in transparent and explainable AI. For researchers, advancing and evaluating AI performance should be complemented by efforts to enhance transparency and explainability of results, a strategy that can nurture deeper trust and broader acceptance in real-world applications.

6 General Discussion

6.1 Summary

The emergence of Generative AI is revolutionizing the marketing landscape through its *dual role*: it serves to both generate [9, 22, 55] and analyze content [36, 38, 41, 54, 66]. This dual capacity makes it a flexible and dynamic tool that can handle a wide range of marketing tasks. Sentiment analysis based on natural language processing stands out as a prevalent use case in marketing, drawing significant attention across application contexts [2, 28, 65]. This research presents a comprehensive benchmark of three state-of-the-art LLMs, i.e., GPT-3.5, GPT-4, and Llama 2, in sentiment analysis, benchmarking their performance with fine-tuned transfer-learning models. Drawing on the expanded empirical framework of Hartmann et al. (2023), we identified and evaluated critical factors that influence the classification accuracy of LLMs in sentiment analysis.

First, we explored the zero-shot sentiment analysis capabilities of state-of-the-art LLMs on binary and three-class classification tasks. This study involved an analysis of over 3,900 unique text documents sourced from 20 different online review datasets. Overall, we found that LLMs are on par or even outperform traditional transfer learning methods in binary and three-class sentiment analysis tasks. GPT-4 showed the strongest performance of all tested LLMs in a zero-shot manner, recording an average accuracy of 93% (binary) and 83% (three-class), surpassing all other models, except for SiEBERT in the binary experiment. Interestingly, Llama 2 performed stronger than GPT-3.5 in the binary classification task with an average accuracy of about 91% despite its smaller size in terms of model parameters (70B vs. 175B). This suggests that parameter count does not solely explain LLMs' performance in sentiment classification. This finding could further drive the open-source movement for LLMs, which is already gaining momentum with the release of powerful new models like Mixtral [34]. Upon examination of the confusion matrices, distinct behavioral patterns were evident among the LLMs. For instance, GPT-3.5 and Llama 2 showed a tendency towards positive interpretations. Conversely, GPT-4 demonstrated a slight bias towards negative interpretations. Surprisingly, all models occasionally deviated from their instructions for the binary classification task ("positive" vs. "negative") and occasionally classified reviews as "neutral".

Second, we evaluated how data characteristics, linguistic features, and analytical procedure affect LLM performance. All models demonstrated lower accuracy on Twitter datasets compared to other user-generated online product reviews, with LLMs particularly challenged by the colloquial and ambiguous language common on Twitter. Content-laden text documents, containing longer words and comprising multi-sentence documents, significantly increased sentiment prediction accuracy across LLMs, highlighting their proficiency with detailed and context-rich content. Interestingly, among the three LLMs, only GPT-3.5 benefited from few-shot and contextual prompting.

Third, we explored the explainability of sentiment classifications generated by LLMs. This study assessed 96 classification explanations generated by LLMs across the 16 datasets from the binary classification experiment, examining their understandability, level of detail, and trustworthiness. All LLMs demonstrated the ability to generate explainable results. Especially Llama 2 showcased an impressive ability to provide understandable and detailed classification explanations, challenging the common perception of AI as inscrutable "black boxes" [53]. Conversely, transfer learning models such as SiEBERT only provide explainability when used in conjunction with other models, such as LIME [21].

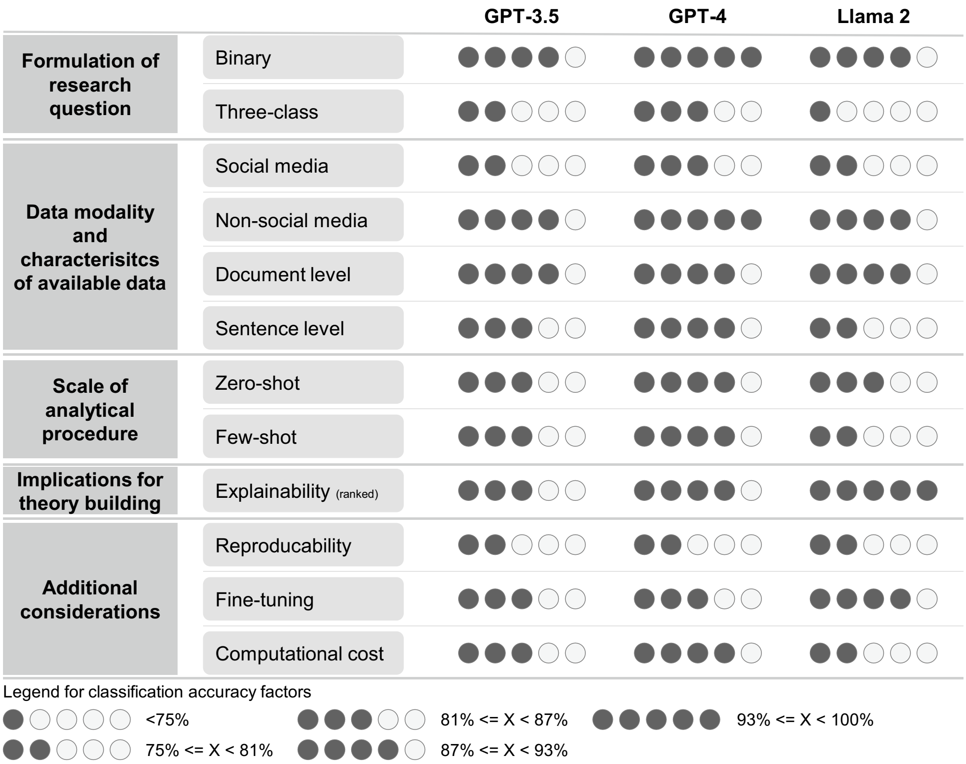
6.2 Implications and Contributions

Our research contributes to the evaluation of Generative AI in sentiment analysis, extending existing research on performance evaluation of LLMs [36, 66]. We offer an extended empirical framework and a multidimensional benchmark, guiding and simplifying method selection for researchers and businesses in the age of Generative AI. Our experimental design builds on a large and diverse data sample, accounting for data contamination and systematic investigation of influencing factors, such as dataset origin, linguistic characteristics, and analytical procedure. For marketing practitioners, the remarkable zero-shot sentiment classification performance achieved by all three tested LLMs underscores a paradigm shift where the convention of developing or fine-tuning models for sentiment analysis on specific and proprietary datasets might become less relevant with the proliferation of state-of-the-art LLMs. Figure 6 summarizes our key findings, serving as a practical guide for sentiment analysis method selection.

Three key caveats are critical to consider when applying LLMs in sentiment analysis: (1) Performance decreases with an increasing number of classes; (2) Data characteristics and analytical procedure significantly influence classification accuracy; (3) Factors such as reproducibility, fine-tuning options, and computing costs also impact method choice.

First, we observe a classification performance drop in LLMs when increasing the number of classes from two to three. Additionally, in a seven-class sentiment classification study on the Google Emotions Dataset [15], the accuracy gap between a fine-tuned transfer learning model (Emotion English DistilRoBERTa-base) [20] and the lowest-performing LLM (Llama 2) was 16.4 percentage points, compared to 8.1 percentage points in the binary experiment (SiEBERT vs. GPT-3.5) (see Web Appendix Table A9 for results). This trend suggests that as classification tasks become more nuanced and specific, the applications of LLMs in a zero-shot setting may be less suitable compared to specifically fine-tuned transfer learning models.

Second, data characteristics, analytical procedure, and inherent tendencies of LLMs matter for the choice of method. It is essential to optimize performance by reducing the model's tendency for certain errors, e.g., confusing negative for positive reviews, which might be problematic if the objective is the detection of social media firestorms [19].



		GPT-3.5	GPT-4	Llama 2
Formulation of research question	Binary	●●●●○	●●●●●	●●●●○
	Three-class	●●○○○	●●●○○	●○○○○
Data modality and charactersitcs of available data	Social media	●●○○○	●●●○○	●●○○○
	Non-social media	●●●●○	●●●●●	●●●●○
	Document level	●●●●○	●●●●○	●●●●○
	Sentence level	●●●○○	●●●●○	●●○○○
Scale of analytical procedure	Zero-shot	●●●○○	●●●●○	●●●○○
	Few-shot	●●●○○	●●●●○	●●○○○
Implications for theory building	Explainability (ranked)	●●●○○	●●●●○	●●●●●
Additional considerations	Reproducability	●●○○○	●●○○○	●●○○○
	Fine-tuning	●●●○○	●●●○○	●●●●○
	Computational cost	●●●○○	●●●●○	●●○○○

Legend for classification accuracy factors

●○○○○ <75%
●●○○○ 75% <= X < 81%
●●●○○ 81% <= X < 87%
●●●●○ 87% <= X < 93%
●●●●● 93% <= X < 100%

Fig. 6 Evaluation matrix for method selection summarizing the consolidated results. *Note: The table records average accuracy per feature without accounting for multiple interactions between dataset characteristics (e.g., number of classes and text origin). Nevertheless, it helps researchers to gauge efficacy and the relative performance of LLMs*

Finally, reproducibility of results is a key concern in the application of LLMs in any research context, as highlighted by Ollion et al. (2024) [48]. Setting the temperature parameter to zero results in model outputs that are only near deterministic. Differentiating model outputs under the same external parameter settings are particularly pronounced in closed-source models, which often undergo model updates that can significantly impact their performance [12]. Our observations also indicate that prompt adjustments can substantially influence the performance of models like GPT-3.5. Addressing this challenge, OpenAI has recently introduced an option to specify a seed parameter [49], aiming to provide greater control and yield mostly deterministic outputs. Despite this advancement, achieving full reproducibility with LLMs remains an elusive goal as of now. The variability in reproducibility is further complicated by the differing approaches of LLM fine-tuning. For example, open-source models like Llama 2 are designed to be fine-tuned by users in local environments, offering greater flexibility. In contrast, OpenAI's closed-source models, such as GPT-3.5 and GPT-4, take a more restrictive fine-tuning approach that requires the submission of fine-tuning data to OpenAI's servers [51]. This restriction could make fine-tuning OpenAI models impractical for any tasks involving proprietary or sensitive data. Moreover, an essential caveat to consider is the significant computational costs associated with advanced models, such as GPT-4 (e.g., output cost per thousand tokens of GPT-4 32 k are 30 times higher compared to GPT-3.5 turbo 16 K [50]). Tools such as the TCO Calculator [60] can help users gauge the total costs of different sentiment analysis options.

In summary, while LLMs present a powerful option for sentiment analysis, careful consideration of model selection, data characteristics, and practical constraints such as reproducibility and cost are essential for optimal application. Besides, ethical considerations should always play a crucial role in the selection of sentiment analysis methods. The use of LLMs in this area requires careful attention to data protection and consent, especially if it involves the processing of extensive user-generated content. This raises significant ethical issues in relation to the protection of personal data and the need to obtain informed consent. In addition, there is a risk that LLMs perpetuate existing biases due to their training on historical datasets. Robust measures are required to ensure fairness and objectivity to prevent outcomes that may treat certain groups or individuals unfairly.

6.3 Limitations and Future Research Directions

Our research, while comprehensive, acknowledges the following limitations which also provide opportunities for future exploration: First, our study was designed to evaluate the off-the-shelf applicability of LLMs in sentiment analysis, meaning we did not fine-tune or specifically train the models for the tasks. It is imperative to investigate opportunities beyond the zero-shot and few-shot capabilities of LLMs.

Hence, the results from our study likely represent the lower bound of the potential of LLMs in sentiment analysis, suggesting that domain-specific training or model fine-tuning can yield even more accurate results. Of special interest is evaluating the influence of training data scale on LLMs' sentiment analysis performance, especially in domains requiring specialized expertise, such as financial or medical advice.

Second, while our research provides insights into the performance differences of LLMs in document level and sentence level sentiment analysis, the area of aspect-based sentiment analysis warrants further investigation. Aspect-based sentiment analysis aims to extract specific aspects or features of the subject of the review (WHAT), classify the sentiment (HOW), and optionally explain the rationale (WHY) [52]. Aspect-based sentiment analysis can help firms understand consumer needs [62] and motivations [11]. Future research could compare the capabilities and examine inherent biases of state-of-the-art LLMs on aspect-based sentiment analysis in a zero-shot and few-shot setting comparable to our experiment.

Finally, building on our study focusing on the data modality of text for sentiment analysis, future research should investigate the performance of Generative AI in different data modalities. The emerging fields of image-based sentiment analysis [69] and multimodal analysis of videos [68], which combine image and audio, offer a wide range of possibilities for deeper and more differentiated sentiment extraction. Additionally, the application of sentiment analysis to real-time data streams such as social media posts [31] and live-streamed comments [13] presents opportunities for leveraging Generative AI in capturing immediate public sentiment. These areas promise to expand the scope and applicability of Generative AI in sentiment analysis, pushing the boundaries of current methodologies and technologies.

7 Conclusion

The emergence of Generative AI tools will have a substantial impact on sentiment analysis research. This paper unveils that LLMs not only compete with but sometimes exceed high-performing transfer learning models in sentiment classification accuracy, making them suitable for immediate business integration. Our analysis extends beyond a performance benchmark and examines how factors like analytical procedure, linguistic features, and data characteristics, such as origin and text length, significantly influence LLM performance. It is essential to manage the use of LLMs in research, considering their inherent biases, tendencies, and reproducibility challenges as well as to ensure that the results are both accurate and ethical. We hope our paper inspires future work on the transformative potential of Generative AI for marketing research in sentiment analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40547-024-00143-4.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Azam W (2022) Headphone Dataset Review Analysis. https://www.kaggle.com/datasets/mdwaquarazam/headphone-dataset-review-analysis. Accessed 24 Aug 2023
2. Berger J, Humphreys A, Ludwig S et al (2020) Uniting the Tribes: Using Text for Marketing Insight. J Mark 84(1):1–25. https://doi.org/10.1177/0022242919873106
3. Berger J, Milkman KL (2012) What Makes Online Content Viral? J Mark Res 49(2):192–205. https://doi.org/10.1509/jmr.10.0353
4. Berger J, Packard G, Boghrati R et al (2022) Marketing insights from text analysis. Mark Lett 33(3):365–377. https://doi.org/10.1007/s11002-022-09635-6
5. Berger J, Sherman G, Ungar L (2020) TextAnalyzer. http://textanalyzer.org/about. Accessed 15 Jan 2024
6. Boyd RL, Ashokkumar A, Seraj S et al (2022) The development and psychometric properties of LIWC-22. University of Texas at Austin, Austin, TX, pp 1–47. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf
7. Brand J, Israeli A, Ngwe D (2023) Using GPT for Market Research. SSRN J. https://doi.org/10.2139/ssrn.4395751
8. Brown T, Mann B, Ryder N et al (2020) Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R et al (eds) Advances in Neural Information Processing Systems, vol 33. Curran Associates Inc, pp 1877–1901
9. Brynjolfsson E, Li D, Raymond L (2023) Generative AI at Work. Natl Bur Econ Res. https://doi.org/10.3386/w31161
10. Castellanos M, Ghosh R, Lu Y et al (2011) LivePulse. In: Sadagopan S, Ramamritham K, Kumar A et al (eds) Proceedings of the 20th international conference companion on World wide web. ACM, New York, NY, USA, pp 193–196
11. Chakraborty I, Kim M, Sudhir K (2022) Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure

and Missing Attributes. J Mark Res 59(3):600–622. https://doi.org/10.1177/00222437211052500

12. Chen L, Zaharia M, Zou J (2023) How is ChatGPT's behavior changing over time? arXiv. https://doi.org/10.48550/arXiv.2307.09009
13. Chouhan A, Halgekar A, Rao A et al (2021) Sentiment Analysis of Twitch.tv Livestream Messages using Machine Learning Methods. In: 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE, pp 1–5
14. Chui M, Yee L, Hall B, Singla A, Sukharevsky A (2023) The state of AI in 2023: Generative AI's breakout year. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year. Accessed 17 Aug 2023
15. Demszky D, Movshovitz-Attias D, Ko J et al. (2020) GoEmotions: A Dataset of Fine-Grained Emotions. arXiv. https://doi.org/10.48550/arXiv.2005.00547
16. Ding N, Qin Y, Yang G et al (2023) Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell 5(3):220–235. https://doi.org/10.1038/s42256-023-00626-4
17. Dwivedi YK, Kshetri N, Hughes L et al (2023) Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manage 71:102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
18. Giray L (2023) Prompt Engineering with ChatGPT: A Guide for Academic Writers. Ann Biomed Eng 51(12):2629–2633. https://doi.org/10.1007/s10439-023-03272-4
19. Hansen N, Kupfer A-K, Hennig-Thurau T (2018) Brand crises in the digital age: The short- and long-term effects of social media firestorms on consumers and brands. Int J Res Mark 35(4):557–574. https://doi.org/10.1016/j.ijresmar.2018.08.001
20. Hartmann J (2022) Emotion English DilstilRoBERTa-base, https://huggingface.co/j-hartmann/emotion-english-distilroberta-base
21. Hartmann J, Bergner A, Hildebrand C (2023) MindMiner: Uncovering linguistic markers of mind perception as a new lens to understand consumer–smart object relationships. J Consum Psychol 33(4):645–667. https://doi.org/10.1002/jcpy.1381
22. Hartmann J, Exner Y, Domdey S (2023) The power of generative marketing: Can generative AI reach human-level visual marketing content? SSRN J. https://doi.org/10.2139/ssrn.4597899
23. Hartmann J, Heitmann M, Schamp C et al (2021) The Power of Brand Selfies. J Mark Res 58(6):1159–1177. https://doi.org/10.1177/00222437211037258
24. Hartmann J, Heitmann M, Siebert C et al (2023) More than a Feeling: Accuracy and Application of Sentiment Analysis. Int J Res Mark 40(1):75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005
25. Hartmann J, Huppertz J, Schamp C et al (2019) Comparing automated text classification methods. Int J Res Mark 36(1):20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009
26. Hartmann J, Netzer O (2023) Natural Language Processing in Marketing. In: Sudhir T (ed) Artificial Intelligence in Marketing. Emerald Publishing Limited, Bingley, pp 191–215
27. Hickey D, Schmitz M, Fessler D et al (2023) Auditing Elon Musk's Impact on Hate Speech and Bots. ICWSM 17:1133–1137. https://doi.org/10.1609/icwsm.v17i1.22222
28. Hirschberg J, Manning CD (2015) Advances in natural language processing. Science 349(6245):261–266. https://doi.org/10.1126/science.aaa8685
29. Hoffman RR, Mueller ST, Klein G et al. (2018) Metrics for Explainable AI: Challenges and Prospects. arXiv. https://doi.org/10.48550/arXiv.1812.04608
30. Homburg C, Ehm L, Artz M (2015) Measuring and Managing Consumer Sentiment in an Online Community Environment. J Mark Res 52(5):629–641. https://doi.org/10.1509/jmr.11.0448
31. Hu A, Flaxman S (2018) Multimodal Sentiment Analysis To Explore the Structure of Emotions. In: Guo Y, Farooq F (eds) Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, New York, NY, USA, pp 350–358
32. Hu EJ, Shen Y, Wallis P et al. (2021) LoRA: Low-Rank Adaptation of Large Language Models. arXiv. https://doi.org/10.48550/arXiv.2106.09685
33. Huang S, Mamidanna S, Jangam S et al. (2023) Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations. arXiv. https://doi.org/10.48550/arXiv.2310.11207
34. Jiang AQ, Sablayrolles A, Roux A et al. (2024) Mixtral of Experts. arXiv. https://doi.org/10.48550/arXiv.2401.04088
35. Kavitha G, Saveen B, Imtiaz N (2018) Discovering Public Opinions by Performing Sentimental Analysis on Real Time Twitter Data. 2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET). IEEE, pp 1–4
36. Kheiri K, Karimi H (2023) SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning. arXiv. http://arxiv.org/pdf/2307.10234v2
37. Kincaid JP, Fishburne J, Robert P. R et al. (1975) Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Defense Technical Information Center, Fort Belvoir, VA
38. Konrad M, Hartmann J (2023) One model fits all? Exploring the zero-shot capabilities of multimodal large language models for automated marketing image analytics. In: Proceeding of the 2023 Marketing Dynamics Conference
39. Larochelle H, Dumitro E, Yoshua B (2008) Zero-Data Learning of New Tasks. AAAI 1(2):646–651
40. Li P, Castelo N, Katona Z et al. (2024) Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. Mark Sci. https://doi.org/10.1287/mksc.2023.0454
41. Marjieh R, Sucholutsky I, van Rijn P et al. (2023) Large language models predict human sensory judgments across six modalities. arXiv. https://arxiv.org/abs/2302.01308
42. Meire M, Hewett K, Ballings M et al (2019) The Role of Marketer-Generated Content in Customer Engagement Marketing. J Mark 83(6):21–42. https://doi.org/10.1177/0022242919873903
43. Meta (2023) LlaMa 2 Model Card. https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md. Accessed 12 Aug 2023
44. Nandwani P, Verma R (2021) A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 11(1):81. https://doi.org/10.1007/s13278-021-00776-6
45. Netzer O, Feldman R, Goldenberg J et al (2012) Mine Your Own Business: Market-Structure Surveillance Through Text Mining. Mark Sci 31(3):521–543. https://doi.org/10.1287/mksc.1120.0713
46. Nguyen N, Johnson J, Tsiros M (2023) Unlimited Testing: Let's Test Your Emails with AI. Mark Sci 0(0). https://doi.org/10.1287/mksc.2021.0126
47. Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. Science 381(6654):187–192. https://doi.org/10.1126/science.adh2586
48. Ollion É, Shen R, Macanovic A et al (2024) The dangers of using proprietary LLMs for research. Nat Mach Intell 6(1):4–5. https://doi.org/10.1038/s42256-023-00783-6
49. OpenAI (2023) Guide to text generation: Reproducible Outputs. https://platform.openai.com/docs/guides/text-generation/reproducible-outputs. Accessed 26 Jan 2024
50. OpenAI (2023) Pricing: Language Models. https://openai.com/pricing. Accessed 24 Aug 2023
51. OpenAI (2023) Guide to fine-tuning: Create a fine-tuned model. https://platform.openai.com/docs/guides/fine-tuning/create-a-fine-tuned-model. Accessed 26 Jan 2024

52. Peng H, Xu L, Bing L et al (2020) Knowing What, How and Why: A Near Complete Solution for Aspect-Based Sentiment Analysis. AAAI 34(05):8600–8607. https://doi.org/10.1609/aaai.v34i05.6383
53. Rai A (2020) Explainable AI: from black box to glass box. J of the Acad Mark Sci 48(1):137–141. https://doi.org/10.1007/s11747-019-00710-5
54. Rathje S, Mirea D-M, Sucholutsky I et al. (2023) GPT is an effective tool for multilingual psychological text analysis. PsyArXiv. https://doi.org/10.31234/osf.io/sekf5
55. Reisenbichler M, Reutterer T, Schweidel DA et al (2022) Frontiers: Supporting Content Marketing with Natural Language Generation. Mark Sci 41(3):441–452. https://doi.org/10.1287/mksc.2022.1354
56. Reiss MV (2023) Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. arXiv. https://doi.org/10.48550/arXiv.2304.11085
57. Reynolds L, McDonell K (2021) Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In: Kitamura Y, Quigley A, Isbister K et al (eds) Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp 1–7
58. Rytting CM, Sorensen T, Argyle L et al. (2023) Towards Coding Social Science Datasets with Language Models. arXiv. https://doi.org/10.48550/arXiv.2306.02177
59. Schneider M (2023) GPT-4 architecture, datasets, costs and more leaked. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/. Accessed 14 Aug 2023
60. Mithril Security (2023) AI TCO Comparison Calculator. https://huggingface.co/spaces/mithril-security/TCO_calculator. Accessed 29 Aug 2023
61. Simmering PF, Huoviala P (2023) Large language models for aspect-based sentiment analysis. arXiv. https://doi.org/10.48550/arXiv.2310.18025
62. Timoshenko A, Hauser JR (2019) Identifying Customer Needs from User-Generated Content. Mark Sci 38(1):1–20. https://doi.org/10.1287/mksc.2018.1123
63. Touvron H, Martin L, Stone K et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv. http://arxiv.org/pdf/2307.09288v2
64. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://doi.org/10.1126/science.aap9559
65. Wang J, Fan Y, Palacios J et al (2022) Global evidence of expressed sentiment alterations during the COVID-19 pandemic. Nat Hum Behav 6(3):349–358. https://doi.org/10.1038/s41562-022-01312-y
66. Wang Z, Xie Q, Ding Z et al. (2023) Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. arXiv. http://arxiv.org/pdf/2304.04339v1
67. Wen Q, Gloor PA, Fronzetti Colladon A et al (2020) Finding top performers through email patterns analysis. J Inf Sci 46(4):508–527. https://doi.org/10.1177/0165551519849519
68. Wu T, Peng J, Zhang W et al (2022) Video sentiment analysis with bimodal information-augmented multi-head attention. Knowl-Based Syst 235:107676. https://doi.org/10.1016/j.knosys.2021.107676
69. You Q, Luo J, Jin H et al (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. AAAI 29(1). https://doi.org/10.1609/aaai.v29i1.9179
70. Zhang Q, Wang W, Chen Y (2020) Frontiers: In-Consumption Social Listening with Moment-to-Moment Unstructured Data: The Case of Movie Appreciation and Live Comments. Mark Sci 39(2):285–295. https://doi.org/10.1287/mksc.2019.1215
71. Zhou Y, Muresanu AI, Han Z et al. (2022) Large Language Models Are Human-Level Prompt Engineers. arXiv. https://doi.org/10.48550/arXiv.2211.01910