Large Language Models Powered Aspect-Based Sentiment Analysis for Enhanced Customer Insights

Análise de Sentimentos por Aspetos com Modelos de Linguagem de Grande Escala para Melhor Compreensão dos Clientes

Mariana Água
NOVA Information Management School (NOVA IMS), Portugal, 20220704@novaims.unl.pt

Nuno António
NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa & Centre for Tourism Research, Development and Innovation (CiTUR), Portugal, nantonio@novaims.unl.pt

Paulo Carrasco
School of Management, Hospitality and Tourism (ESGHT), Universidade do Algarve & Centre for Tourism Research, Development and Innovation (CiTUR), Portugal, pcarras@ualg.pt

Carimo Rassal
School of Management, Hospitality and Tourism (ESGHT), Universidade do Algarve, & CIDEHUS - Interdisciplinary Center for History, Cultures, and Societies of the University of Évora, Portugal, chrassal@icloud.com

Received: 24.09.2024; Revisions required: 09.11.2024; Accepted: 09.12.2024

Abstract

In the age of social networks, user-generated content has become vital for organizations in tourism and hospitality. Traditional sentiment analysis methods often struggle to process large volumes of data and capture implicit sentiments. This study examines the potential of Aspect-Based Sentiment Analysis (ABSA) using Large Language Models (LLMs) to enhance sentiment analysis. By employing GPT-4o via ChatGPT, we benchmark three approaches: a fuzzy logic-based method, manual human analysis, and a new ChatGPT-based analysis. We analyze a dataset of 500 all-inclusive hotel reviews, comparing these methods to assess ChatGPT's effectiveness in identifying nuanced language and handling subjectivity. The findings reveal a high similarity between ChatGPT and human analysis, showcasing ChatGPT's ability to interpret complex sentiments and automate sentiment classification tasks. This study highlights the potential of LLMs in transforming customer feedback analysis, providing deeper insights, and improving responsiveness in the hospitality industry. These results contribute to academia by presenting a framework for using LLMs in ABSA and guiding future applications and development.

Keywords: Automated Sentiment Analysis; Aspect-Based Sentiment Analysis; Large Language Models; Customer Feedback Analysis; ChatGPT Applications; Natural Language Processing.

Resumo

Na era das redes sociais, o conteúdo gerado por utilizadores tornou-se essencial para organizações de turismo e hospitalidade. Métodos tradicionais de análise de sentimentos têm dificuldade em processar grandes volumes de dados e captar sentimentos implícitos. Este estudo explora o potencial da Análise de Sentimentos Baseada em Aspetos (ABSA) com Modelos de Linguagem de Grande Escala (LLMs) para melhorar a análise de sentimentos. Utilizando o GPT-4o via ChatGPT, são comparados três métodos: lógica difusa, análise manual e análise baseada no ChatGPT. Avaliaram-se 500 avaliações de hotéis all-inclusive, analisando a eficácia do ChatGPT na identificação de nuances linguísticas e subjetividade. Os resultados mostram elevada concordância entre ChatGPT e análise humana, destacando a capacidade do modelo em interpretar sentimentos complexos e automatizar classificações. Este estudo evidencia o potencial dos LLMs para transformar a análise de feedback, oferecendo insights mais profundos e melhorando a capacidade de resposta. Contribui para a academia com um modelo de aplicação de LLMs em ABSA, orientando avanços futuros.

Palavras-chave: Análise Automatizada de Sentimentos; Análise de Sentimentos Baseada em Aspetos; Modelos de Linguagem de Grande Escala; Análise de Feedback de Clientes; Aplicações do ChatGPT; Processamento de Linguagem Natural.

1. Introduction

Customers' opinions are paramount in today's business landscape, particularly in the hospitality industry. As companies deepen their understanding of consumer preferences and feedback's influence, this understanding becomes indispensable for ensuring business success (Sudirjo et al., 2023).

Before the 2000s, linguistics and natural language processing research rarely explored opinions and sentiments (Devika et al., 2016). However, the rise of digital platforms and the growth of online communication transformed this scenery. With the advent of the internet, especially social media and review forums (Hussein, 2018), customers have gained new ways to share their experiences, creating a vast volume of data that companies can analyze to obtain valuable insights. Applying artificial intelligence (AI) in this context through Natural Language Processing (NLP) techniques becomes crucial. NLP is a field of computer science that focuses on the interaction between computers and human language and seeks to improve the understanding and processing of human language, applying itself in areas such as language recognition, translation, and text summarization. These techniques offer meaningful insights from a vast data set, providing a deeper understanding of consumer preferences, opinions, and needs (Korkmaz et al., 2023).

Sentiment analysis, or opinion mining, is an area within NLP that automatically identifies a text's sentiment (Hoang et al., 2019) and categorizes it as positive, negative, or neutral. In its early stages, sentiment analysis was dominated by lexicon-based techniques, which involved calculating the number of positive and negative words in a text, with each word associated with a sentiment score (Oliveira et al.,2022; Carvalho et al., 2024; Kheiri & Karimi, 2023). Advances in Machine Learning (ML) algorithms have proved more effective than the techniques implemented until now. However, due to the complexity and nuances of human language, it has been necessary to refine these traditional techniques.

Sentiment analysis has been studied at various levels, including document, paragraph, sentence, and aspect (Chifu & Fournier, 2023; Wankhade et al., 2022). Initially, this analysis focused on the binary sentiment classification, usually positive or negative. However, this approach became insufficient as it was recognized that a single evaluation could express multiple feelings on different subjects, containing opinions with opposing polarities (Kontonatsios et al., 2023)

To overcome this limitation, Aspect-Based Sentiment Analysis (ABSA) emerged. This approach classifies sentiment and identifies the specific aspects to which the evaluation refers. Other methodologies, such as topic modeling, play a complementary role in extracting insights from textual data. Topic modeling, a widely used technique in natural language processing, identifies underlying themes or clusters in large datasets, providing a broader perspective on customer feedback. Studies like Aguilar-Moreno et al. (2024) have demonstrated the effectiveness of topic modeling in supporting business decision-making by categorizing unstructured data into actionable themes. While this study focuses on ABSA using ChatGPT to classify sentiments at a granular level, integrating topic modeling in future research could uncover overarching patterns, adding strategic value to sentiment analysis workflows. This study builds on ABSA by leveraging ChatGPT to address gaps in sentiment analysis, enhancing precision in handling subjectivity and nuanced language. Doing so represents an evolution in the field, enabling the extraction of more detailed information about each opinion compared to traditional approaches (Zhang et al., 2023).

ABSA addresses two main tasks: aspect detection and sentiment classification. In the aspect detection task, the aim is to identify the entities or characteristics of the text that are the subject of an opinion. This detection can be explicit, involving the analysis of linguistic patterns, or implicit, requiring a deeper comprehension. Once the aspects have been identified, the sentiment is given a classification. This approach improves the accuracy of sentiment analysis and offers more detailed insights into consumer perceptions of specific aspects (Truşcă, et al., 2023).

In the last decade, the field of NLP has experienced a paradigm shift due to exponential data growth and a revolution driven by remarkable advances in AI. With the rise of Large Language Models (LLMs), especially the powerful Generative Pretrained Transformer (GPT), sentiment analysis achieved substantial improvements, promoting significant advances in the automated understanding of emotions in texts (Liu et al., 2023). As LLMs gain prominence, their ability to understand and generate human language on a large scale emerges. These models are trained with large amounts of text data, allowing them to learn complex patterns and relationships in human language (Naveed et al., 2023). The remarkable development of LLMs is primarily attributed to OpenAI, which released the first GPT algorithm in 2018 (Tan et al., 2023). Since then, several versions of GPT have been released, with GPT-4o being the most recent model.

This research is driven by a continuous search for innovation and improvement in business strategy. It recognizes the need to simplify and optimize the analysis of customer reviews. This study aims to provide organizations with an efficient method for identifying overall customer sentiment and the specific aspects influencing those sentiments. To achieve this goal, the GPT-4o algorithm, through ChatGPT, will be used to automate specific tasks, such as identifying sentiments and aspects in customer reviews.

Although recent studies have shown success in classifying customer reviews using LLMs (Simmering & Huoviala, 2023), there needs to be more literature addressing the need for the use of more advanced models. To overcome this gap, this study proposes an approach using ChatGPT with a dataset of TripAdvisor hotel reviews (Rassal et al., 2023). By analyzing the performance of ChatGPT, the goal is to address gaps identified in existing literature and propose solutions for enhancing sentiment analysis and aspect identification in customer reviews. To guide this investigation, two research questions have been formulated to focus on the study's objectives:

RQ1: How can sentiment analysis leveraging Large Language Models (LLMs) enhance the capabilities of Aspect-Based Sentiment Analysis (ABSA)?

RQ2: How does the performance of an LLM-based approach to ABSA compare to a reference human analysis approach?

The methodology employed in this research combines quantitative methods with the evaluation of hotel online reviews using various approaches using the ChatGPT API (API Reference - OpenAI API, n.d.). A comprehensive literature review will be conducted to establish the theoretical foundations, focusing on aspect-based sentiment analysis. Subsequently, the ChatGPT API, using the GPT-4o model, is employed to identify aspects and categorize sentiment in reviews accurately. Following this, a detailed evaluation is conducted, employing the Similarity Coefficient within a benchmarking framework to compare three distinct approaches: a Fuzzy logic-based method (Rassal et al., 2023), the ChatGPT approach, and a Human approach.

The remainder of the paper is organized as follows. Section 2 reviews the literature, tracing the evolution of sentiment analysis from traditional methods to Aspect-Based Sentiment Analysis (ABSA), emphasizing the transformative role of Large Language Models (LLMs). Section 3 details the methodology, including dataset preparation, the design of benchmarking frameworks, and the comparative evaluation of ChatGPT's performance against human analysis and traditional fuzzy logic-based approaches. Section 4 presents the results, offering a comprehensive analysis of findings across dimensions. Section 5 expands upon these findings with a detailed discussion, situating ChatGPT's performance within the context of existing literature, identifying alignment and disparities with prior research, and highlighting its key contributions while discussing theoretical and practical implications. Finally, Section 6 concludes the paper by summarizing the study's findings, acknowledging limitations, and presenting directions for future research.

2. Literature review

2.1 *Aspect-Based Sentıment Analysıs*

Sentiment analysis has been extensively studied across various levels (Wankhade et al., 2022), starting with analyzing entire documents. Initially, researchers focus on determining the overall sentiment of a document, aiming to discern whether it expresses positivity or negativity, commonly employing conventional machine learning methods (Kumar & Sebastian, 2012; Khalaf et al., 2022). As the analysis progresses to a more granular level, namely the sentence level, the focus shifts towards identifying opinions within individual sentences, where the goal is to differentiate between positive, negative, and neutral sentiments.

With the exponential increase in sentiment analysis, the ABSA approach, proposed in 2010 as a new framework (Thet et al., 2010), brought a more refined perspective, allowing the assessment of specific sentiments about different aspects of an entity (Wang & Liu, 2022). Further illustrating the utility of sentiment analysis in practical contexts, Aguilar-Moreno et al. (2024) highlighted its increasing role in business decision-making, integrating multicriteria decision-making and predictive algorithms to process customer feedback efficiently. These approaches address the need for refined sentiment analysis techniques to explore specific aspects and their implications. They have become essential for uncovering nuances present in various evaluations, enabling a more detailed analysis of specific aspects (Serrano-Guerrero et al., 2015).

Exploring the review "The device is expensive, but the camera is amazing," it is possible to identify an explicit aspect, "camera" (sentiment term: "amazing", polarity: positive), and an implicit aspect, "price" (sentiment term: "expensive", polarity: negative). Notably, depending on the areas being considered, these aspects can be categorized into broader categories, such as "functionality" for "camera" and "cost" for "price."

As articulated by Liu (2012), ABSA relies on four key elements: aspect category (c), aspect term (a), opinion term (o), and sentiment polarity (p): The aspect term (a) is the target of the opinion that appears explicitly in the text provided, such as "device" in the sentence "The device is expensive". When the target is not directly indicated in the sentence (for example, "It's over budget!"), it can be considered "null" or absent. The aspect category (c) refers to the broader field or domain to which an aspect belongs. They can be predefined or extracted from the text and play a crucial role in contextualizing sentiment analysis results. Taking the previous example, the category "design" or "price" can be associated with the aspect "device", while the category "performance" or "quality" can be attributed to the aspect "camera". Opinion terms (o) can be expressed explicitly or implicitly. In an explicit expression, such as presented, the positive attitude towards the camera is clearly indicated, while the attitude towards the price is implicit in the word "expensive". Sentiment polarity (p) describes the orientation of sentiment towards a spectrum category or term, usually categorized as positive, negative, or neutral. In this context, sentiment polarity could be negative due to the expression "expensive".

2.2 *Related Work*

ABSA has evolved significantly from its earliest stages to recent developments, encompassing various techniques and approaches. Initially, manual methods based on linguistic rules, such as those introduced by Hu and Liu (2004) and Pang and Lee (2004), were pioneers in this area.

While Hu and Liu (2004) opted to categorize aspects and sentiments in a lexicon, Pang and Lee (2004) explored an innovative method based on graph theory to identify and summarise subjective portions of text. However, both methods faced challenges, such as dealing with implicit aspects and the varied interpretations of sentiment.

Over time, the emergence of academic competitions, such as SemEval, stimulated innovation in ABSA. Studies conducted by Pontik et al. (2014, 2015, 2016) were crucial in this process, introducing more sophisticated approaches that addressed issues such as subjectivity detection, aspect extraction, and polarity classification in a more precise and comprehensive way. However, even with these advances, persistent challenges, such as the representativeness of the data and the generalisability of the results to different domains, persisted.

Al-Smadi et al. (2019) marked the shift to automated approaches using machine learning and neural networks, showing their effectiveness in handling data complexity and improving sentiment analysis accuracy. However, despite the notable improvements achieved by neural network models, concerns persist about their interpretability and applicability in different contexts (Rathan et al., 2018; Pham & Le 2018).

Rathan et al. (2018) proposed a model for analyzing sentiment in smartphone reviews that employ methods such as supervised learning with decision trees and Support Vector Machines (SVM). Although this contributes to classification accuracy, limitations arise due to the presence of ambiguities in the evaluations, which can affect the model's effectiveness in real-world situations. On the other hand, the model proposed by Pham and Le (2018) emphasizes the ability to represent semantic relationships in several layers. However, its specificity to the hospitality sector may limit its generalisability to other domains. In addition, the model needs to fully address implicit aspects of evaluations, causing difficulties in interpreting ambiguous or sarcastic expressions.

Mojica et al. (2024) explored the application of sentiment analysis in happiness apps, focusing on user engagement and well-being. Their findings highlighted the role of gamification in enhancing user interaction but also emphasized the methodological limitations, such as reliance on user reviews and limited clinical validation. Perea-Khalifi et al. (2024) applied sentiment analysis to P2P payment systems, identifying six latent aspects influencing user experiences, such as ease of use and perceived value. They noted that independent apps evoke more positive emotions than those associated with financial institutions, underscoring the importance of context-specific sentiment analysis.

Recent advances, such as the Transformer architecture and LLMs, have further boosted ABSA's effectiveness. The emergence of transformer models, such as GPT-3, has been essential in meeting the challenges encountered until then, offering capabilities in natural language understanding. The works by Simmering and Huoviala (2023) and Machálíková (2023) highlight the potential of LLMs by demonstrating their superiority over traditional methods, especially in large datasets and more refined sentiment analysis. However, the need for more comprehensive evaluations with diverse datasets to prove the effectiveness of these algorithms remains a point of attention.

2.3 *Large Language Models*

LLMs are advanced language models based on AI, falling into the category of Generative AI (Dasgupta et al., 2023). The architecture underlying the current generation of LLMs is based on transformers, a type of neural network that enables the capture of complex contexts and semantic relationships (Azam et al., 2023).

With an encoder-decoder mechanism and self-attention implementation (Vaswani et al., 2017), these models break down the text into more meaningful units, assigning dynamic weights to each term. This approach enables deep contextual understanding, and by applying this mechanism to tasks such as ABSA, LLMs can pick up on specific nuances about the aspects mentioned (Mishev et al., 2020). Considering the previously mentioned sentence, the device is expensive, but the camera is amazing, it is possible to observe how the architecture of the LLM operates. Using the self-attention mechanism, the encoder divides the sentence into meaningful units, assigning contextual weights to each term. This allows the model to identify two specific aspects: the device price and the camera performance, each with weighted relevance. During decoding, the model generates an output that classifies the sentiments associated with each aspect. Thus, even with the mention of the high cost, emphasizing the amazing camera may lead to a positive classification for this aspect. This dynamic interaction between encoder and decoder, mediated by the self-attention mechanism, allows LLMs to understand and evaluate sentiments in specific aspects.

2.4 *Large Language Models In ABSA*

LLMs stand out in ABSA due to their deep contextual understanding, acquired through extensive training on textual data (Liu et al., 2023). This capability enables accurate sentiment analysis of specific aspects in different contexts, from product reviews to social media posts (Azam et al., 2023). The multilingual ability of LLMs is also valuable for ABSA in global markets, where opinions can vary culturally. In addition, these models are adaptable to new domains, allowing for personalized sentiment analysis. They also stand out for their efficiency, avoiding the need for extensive readiness engineering. These models offer a robust and effective approach, enabling accurate aspect identification and efficient sentiment classification (Zhang et al., 2023).

LLMs have been extensively studied in ABSA, showcasing their versatility across various domains. In customer comment analysis on Twitter, a transfer-based ABSA model (Banjar et al., 2021) stands out in aspect polarity estimation. Additionally, transformer-based approaches like BERT and Transformer demonstrate notable advantages in government (Areed et al., 2020) and financial sectors (Mishev et al., 2020). Addition studies explore LLMs in sentiment analysis across social media (Hoang et al., 2019), restaurant reviews (Li et al., 2023), products (Ismet et al., 2022), and movies (Nkhata , 2022). Table 1 concisely summarizes the studies, including their main conclusions and limitations.

Table 1 - Overview of Studies and Results

Authors (year)	Title	Theme	Main Conclusions	Limitations
Hu & Liu (2004).	Mining and summarizing customer reviews	Sentiment extraction from customer reviews	• Lexicon-based approach to classify customer opinions in relation to different product or service characteristics.	• Difficulty in dealing with implicit aspects. • Problems in dealing with feelings about subjective aspects.
Pang & Lee (2004)	A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts	Sentiment extraction from customer reviews	• Approach based on minimum cut algorithms to identify and summarize subjective portions of text.	• Uncertainty about the effectiveness of the graph-based approach in large databases and with different types of text. • Obstacles due to the varied interpretation of feelings
Pontiki et al. (2014)	SemEval-2014 Task 4: Aspect-Based Sentiment Analysis	Sentiment extraction from customer reviews	• Emphasis on subjectivity detection, aspect extraction, and polarity classification.	
Pontiki et al. (2015)	SemEval-2015 Task 12: Aspect-Based Sentiment Analysis	Sentiment extraction from customer reviews	• Expansion of the scope to consider multiple opinions on the same aspect and cover various domains.	
Pontiki et al. (2016)	SemEval-2016 Task 5: Aspect-Based Sentiment Analysis	Sentiment extraction from customer reviews	• Emphasis on considering cultural and linguistic contexts.	
Al-Smadi et al. (2019)	Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews	Aspect-based sentiment analysis of hotel reviews in Arabic	• The use of bidirectional LSTMs and aspect-based LSTMs led to significant improvements compared to the baseline research.	• Limited to hotel reviews in Arabic.
Rathan et al. (2018)	Consumer Insight Mining: Aspect Based Twitter Opinion Mining of Mobile Phone Reviews	Aspect-based sentiment analysis of smartphone reviews	• The use of decision trees for aspect extraction and SVM for sentiment classification yielded promising results. Additionally, a domain-specific lexicon was employed to enhance classification accuracy.	• The dataset size and ambiguous tweets may limit the model's accuracy.
Pham & Le (2018)	Learning multiple layers of knowledge representation for aspect based sentiment analysis.	Aspect-based sentiment analysis of hotel reviews	• The use of an innovative neural network architecture with multiple layers of knowledge representation led to superior results compared to traditional methods. • The interpretation of ambiguous or sarcastic expressions is not fully addressed.	• The model is specific to the hotel industry and is not generalizable to other domains.
Simmering & Huoviala (2023)	Large language models for aspect-based sentiment analysis.	Aspect-based sentiment analysis	• LLMs such as GPT3 have demonstrated superiority over traditional methods in identifying sentiments related to specific aspects.	• The analysis was restricted to a single dataset.
Macháliková (2023)	Utilizing ChatGPT for Sentiment Analysis Title Utilizing ChatGPT for Sentiment Analysis.	Sentiment analysis based on online reviews	• ChatGPT can be effective in sentiment analysis of online reviews, both for individual reviews and multiple reviews.	• The study was limited to a small dataset and did not compare ChatGPT's performance with other LLMs.
Aguilar-Moreno et al. (2024)	Sentiment analysis to support business decision-making. A bibliometric study.	Sentiment analysis in business decisions	• The study emphasizes the increasing integration of sentiment analysis methods in business decision-making, highlighting their application across various sectors and the importance of customer feedback from platforms like Twitter and Amazon	• Focused on English-language studies and limited consideration of non-Western contexts and lesser-explored sectors.
Mojica et al. (2024)	Is there innovation management of emotions or just the commodification of happiness? A sentiment analysis of happiness apps.	Emotions in happiness apps through sentiment analysis.	•Apps improve well-being and gamification aids engagement.	• Relies on user review and lacks clinical data validation.
Perea-Khalifi et al. (2024)	Exploring the determinants of the user experience in P2P payment systems in Spain: A text mining approach.	User experience in P2P payment apps	• Identified six key aspects influencing user reviews. • Independent apps evoke more positive emotions than bank-affiliated ones.	• Bias toward positive reviews.

Source: Own elaboration.

2.5 *Generatıve Pre-Transformer*

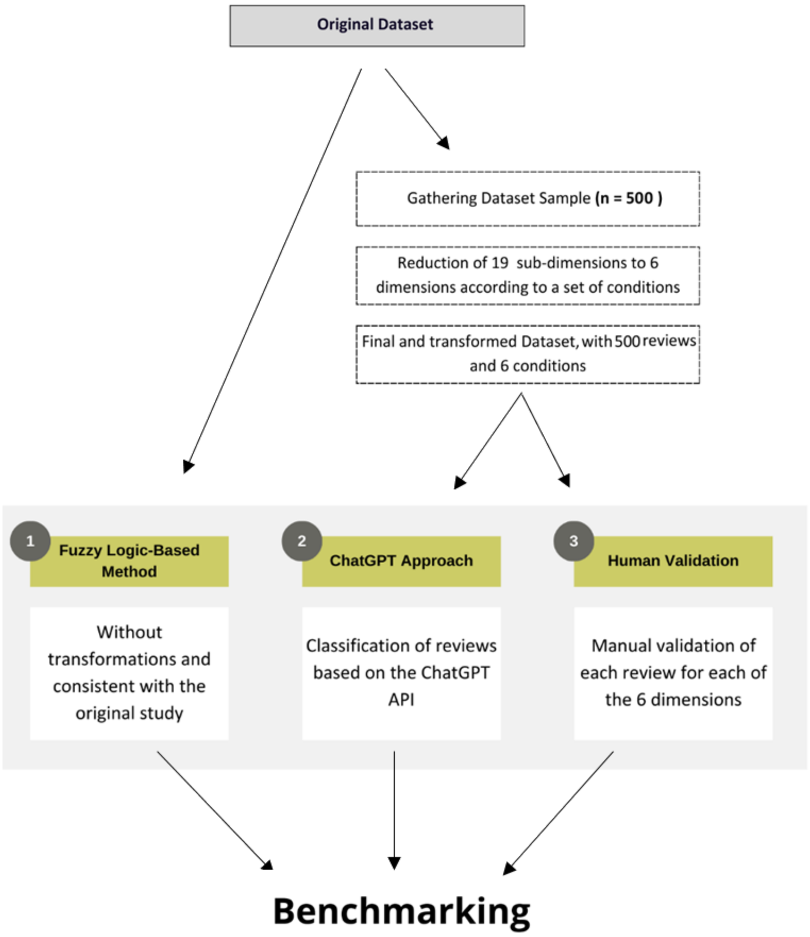
Following its introduction by OpenAI in 2018, one of the most widely adopted LLMs is the GPT (Raj et al., 2023). Derived from the GPT model, ChatGPT (Rudolph et al., 2023) has emerged as a versatile AI model in NLP (Amar, J. et al., 2023). It leverages the Transformer architecture to generate humanized responses in real-time conversations. Trained on vast sets of textual data, this model uses advanced deep-learning techniques to understand and generate coherent text. The model has been trained on millions of online conversations, obtaining relevant information on topic contexts.

Regarding sentiment analysis, ChatGPT has made a name for itself (Baker & Utku, 2023), offering significant advantages for businesses and organizations, enabling effective automation in understanding customer perceptions and feedback. Its ability to conduct contextual analysis contributes to a better understanding of text nuances, enabling more accurate decisions and proactive identification of trends, problems, or opportunities. By automating the analysis of large data sets, ChatGPT can provide insights into community needs and concerns, aiding informed decision-making, identifying priority intervention areas, and adapting strategies to meet community demands better (Sudirjo et al., 2023). Alongside sentiment analysis, customer service is a critical aspect for organizations such as Destination Management Organizations and hotels. ChatGPT can play a pivotal role in this field. By offering virtual support, it delivers exceptional customer service experiences by promptly addressing frequently asked questions, providing detailed information about products and services, and solving everyday problems, which reduces the need for human intervention (George et al., 2023; Yñiguez-Ovando et al. 2024).

3. Methodology

This study utilizes the dataset and builds upon the findings of Rassal et al. (2023) to benchmark the performance of Large Language Models (LLMs) in enhancing Aspect-Based Sentiment Analysis (ABSA). Specifically, it employs GPT-4o, the latest iteration of OpenAI's Generative Pre-trained Transformer, accessed via the OpenAI API. Figure 1 represents this study design.

Figure 1 - Schema of the design of the study


Original Dataset
Gathering Dataset Sample (n = 500)
Reduction of 19 sub-dimensions to 6 dimensions according to a set of conditions
Final and transformed Dataset, with 500 reviews and 6 conditions
1
Fuzzy Logic-Based Method
2
ChatGPT Approach
3
Human Validation
Without transformations and consistent with the original study
Classification of reviews based on the ChatGPT API
Manual validation of each review for each of the 6 dimensions
Benchmarking

Source: Own elaboration.

The dataset used in this study originates from the work of Rassal et al. (2023), which comprises 6,742 validated reviews from four— and five-star all-inclusive hotels from Tripadvisor, a primary platform for customer reviews of accommodation services. These reviews were gathered between August 2003 and April 2019 and encompass 19 sub-dimensions.

Given the costs associated with using ChatGPT and the processing time required (approximately 2 to 4 seconds per review), we opted to analyze a random sample of 500 reviews (n=500) to ensure a more comprehensive representation of the dataset while maintaining feasibility. Rassal et al. (2023) applied a fuzzy sets approach, allowing the assignment of degrees of pertinence to customer descriptions.

The dataset was prepared to ensure compliance with ethical data usage standards and to optimize it for analysis. Large Language Models (LLMs), such as GPT-4o, are inherently trained on vast amounts of real-world text, which enables them to interpret and detect nuanced meanings, even from raw or unprocessed data. This capability allows for a context-aware analysis that can uncover subtle patterns in sentiment. Nevertheless, to enhance the accuracy of the analysis and refine the dataset for research purposes, several preprocessing steps were undertaken. First, all personal identifiers were removed to ensure anonymity and align with ethical guidelines. Each review was assigned a unique code to maintain traceability while safeguarding user privacy. Next, the text underwent noise removal, where non-alphanumeric characters, control characters, URLs, and extraneous symbols were stripped from the data to produce a cleaner input for analysis. Additionally, all text was standardized by converting it to lowercase to ensure uniformity across the dataset.

Given the dataset's initial complexity of 19 sub-dimensions, it was necessary to consolidate it into six dimensions, as the study by Rassal et al. (2023), to ensure an equal representation of various aspects of the reviews. These aspects are Staff, Price, Place, Ambience, Experiences, and Services (see Table 2 for a summary of the dimensions and their corresponding sub-dimensions).

Table 2 - Summarization of Sub-Dimensions into Dimensions

Dimension	Sub-dimensions
Staff	Staff, Language, Empathy, Assurance, Responsiveness and Reliability
Price	Rates, Promotions and Price
Place	Location and Room
Ambience	Hotel and Design
Experiences	Organization and Experience
Services	Facilities, Services, Food and Beverage

Source: Own elaboration.

The Staff dimension encompasses various aspects related to interactions with hotel staff, including their behavior, professionalism (Staff), language proficiency, and communication skills (Language). It also covers the ability of staff to understand and address customers' needs (Empathy), their confidence and knowledgeability (Assurance), responsiveness to requests (Responsiveness), and the consistency of service delivery (Reliability).

The Price evaluates the cost associated with the hotel services, including the standard rates charged for accommodations and amenities (Rates), any promotional offers or discounts available (Promotions), and the overall perceived value paid by customers (Price).

The Place dimension encompasses the hotel's location and room characteristics. Location includes aspects such as proximity to attractions, accessibility, and surroundings (Location). Room evaluates the quality of the accommodations, including cleanliness, comfort, view, and size (Room).

Ambience considers the overall atmosphere and design of the hotel. It encompasses factors such as the general ambiance, mood, and vibe, including the friendliness and warmth of the environment (Hotel). Design evaluates the aesthetics, layout, decoration, and style of the hotel's public and private areas, including rooms, the lobby, and common spaces (Design).

Experiences focus on customers' satisfaction with the hotel's atmosphere and the general quality of the setting. This includes the subjective experience (Experience) during their stay, such as comfort and the pleasantness of the environment, as well as the layout and quality of visual, auditory, and functional elements (Organization). Additionally, it assesses the efficiency and structure of the hotel's service system, aiming to reduce inefficiencies and enhance overall satisfaction by considering waiting time, service efficiency, and satisfaction with the hotel's operations.

The Services dimension assesses the hotel's availability and quality of services. (Facilities) evaluate the amenities, infrastructure, and physical features the hotel provides, including recreational areas, business facilities, and parking. Services examine the efficiency, responsiveness, and quality of customer service provided by staff, including housekeeping, concierge, and front desk assistance (Services). (Food and Beverage) assesses the variety, quality, and presentation of food and beverage options available at the hotel.

Sub-dimensions are evaluated on a scale ranging from -1, 0, 1, and blank, representing different meanings of customers' experiences. The value -1 indicates a negative evaluation, reflecting an unfavorable experience. A rating of 0 indicates a neutral evaluation, where the user does not express a strong opinion, positive or negative. On the other hand, a value of 1 represents a positive evaluation, demonstrating user satisfaction. When the review does not refer to the analyzed sub-dimension, it is classified as blank with no associated sentiment.

Once these values are established for the sub-dimensions, a set of rules must be established to condense them into six dimensions. This step is crucial as it will enable a comparative analysis in the future. These rules are applied separately to each dimension by evaluating the various combinations of values assigned to the sub-dimensions. This process determines the overall sentiment for each dimension, ensuring that the aggregate sentiment reflects the nuances captured within the sub-dimensions. Table 3 describes all the conditions considered.

Table 3 - Conditions employed for aggregating Sub-dimensions

Condition	Overall Sentiment
The majority of sub dimensions rated as 1	1
The majority of sub dimensions rated as -1	-1
The majority of sub dimensions rated as 0	0
All of sub dimensions rated as blank	Blank
Half of sub dimensions rated as 1 and half as -1	0
Half of sub dimensions rated as 1 and half as blank	1
Half of sub dimensions rated as -1 and half as blank	-1

Source: Own elaboration.

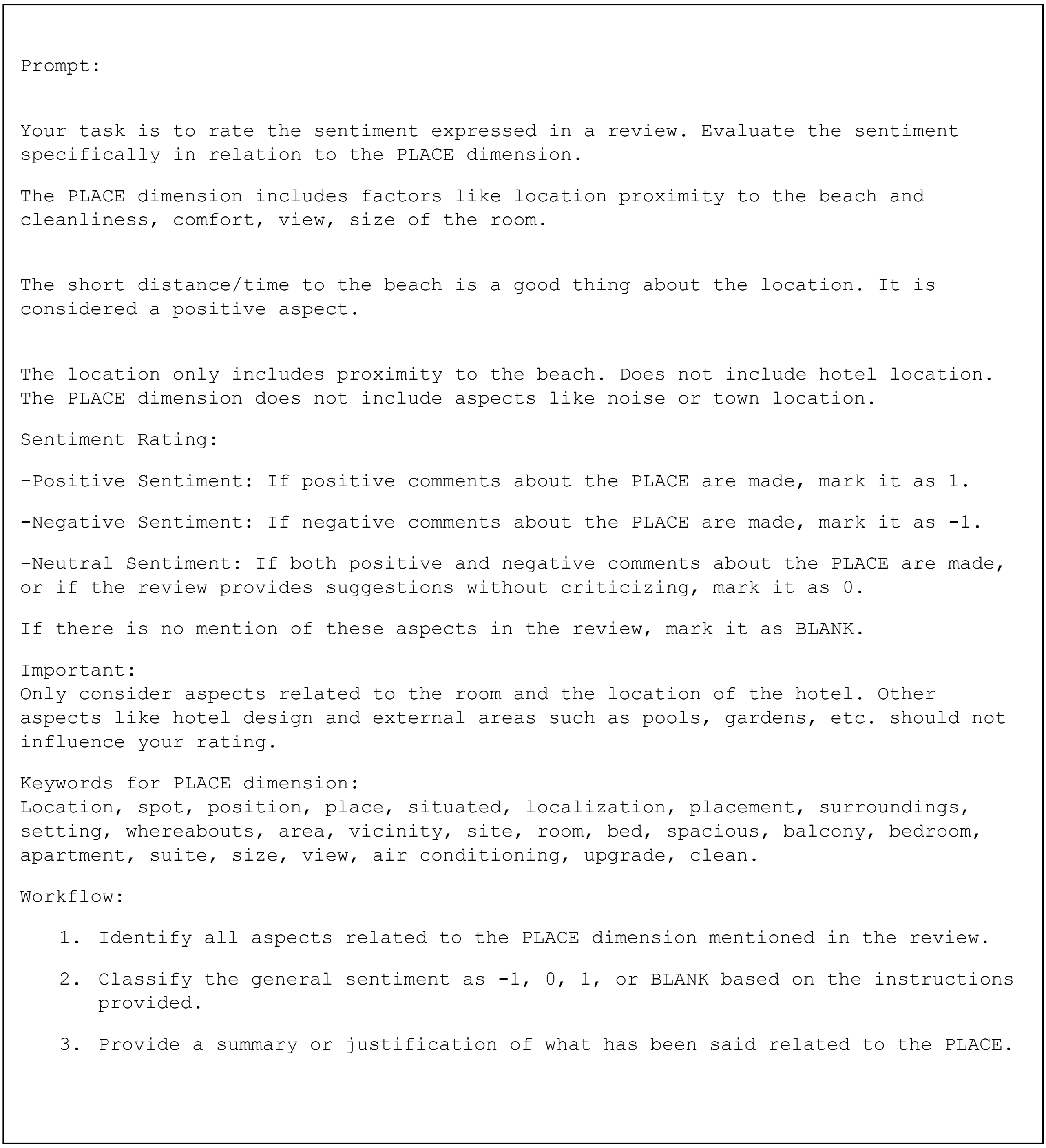
After obtaining the dataset structured into six dimensions and incorporating the results from the previous study, which will serve as the first benchmark to assess the efficacy of the approach with ChatGPT, the next step is to use a ChatGPT approach to extract aspects and analyze sentiments expressed in the reviews.

To enable ChatGPT's analysis, it is necessary to use its Application Programming Interface (API). Specifically, the OpenAI API allows users to send requests to the ChatGPT model hosted on servers maintained by OpenAI. These requests typically consist of textual inputs, such as prompts or questions, which the model processes and responds to accordingly. The API handles the communication between the user's application and the ChatGPT model, providing a scalable and user-friendly infrastructure for deploying AI solutions. Detailed information is available in the OpenAI documentation (OpenAI API, n.d.).

The study utilized the GPT-4o model, chosen for its advanced natural language processing capacity and its ability to generate human-like text. Python programming language and the openai-python library were used to interact with the model. Parameters were carefully chosen to optimize performance: a temperature of 0 was set to ensure deterministic outputs, maximizing consistency in the responses; the maximum token limit was set to 4090 to accommodate longer input sequences; and the top_p parameter was also set to 1, allowing the model to consider all possible tokens when generating responses.

For ChatGPT to classify the reviews accurately, a prompt, which is the message parameter, is necessary for each of the six dimensions. This prompt provides information about each one and examples, ensuring the model clearly understands what to look for and how to assess the sentiment expressed in the reviews. Following this, the objective is for ChatGPT to analyze the sample of reviews and classify them into -1, 1, 0, or blank per dimension. Figure 2 exemplifies the prompt used for the Place dimension.

Figure 2 - Place Custom Prompt



```
Prompt:

Your task is to rate the sentiment expressed in a review. Evaluate the sentiment
specifically in relation to the PLACE dimension.

The PLACE dimension includes factors like location proximity to the beach and
cleanliness, comfort, view, size of the room.

The short distance/time to the beach is a good thing about the location. It is
considered a positive aspect.

The location only includes proximity to the beach. Does not include hotel location.
The PLACE dimension does not include aspects like noise or town location.

Sentiment Rating:

-Positive Sentiment: If positive comments about the PLACE are made, mark it as 1.

-Negative Sentiment: If negative comments about the PLACE are made, mark it as -1.

-Neutral Sentiment: If both positive and negative comments about the PLACE are made,
or if the review provides suggestions without criticizing, mark it as 0.

If there is no mention of these aspects in the review, mark it as BLANK.

Important:
Only consider aspects related to the room and the location of the hotel. Other
aspects like hotel design and external areas such as pools, gardens, etc. should not
influence your rating.

Keywords for PLACE dimension:
Location, spot, position, place, situated, localization, placement, surroundings,
setting, whereabouts, area, vicinity, site, room, bed, spacious, balcony, bedroom,
apartment, suite, size, view, air conditioning, upgrade, clean.

Workflow:

   1. Identify all aspects related to the PLACE dimension mentioned in the review.

   2. Classify the general sentiment as -1, 0, 1, or BLANK based on the instructions
      provided.

   3. Provide a summary or justification of what has been said related to the PLACE.
```

Source: Own elaboration.

To replicate Rassal et al.'s (2023) study and use their results as a benchmark, the dictionary created by those authors was employed to identify the specific aspects addressed in each dimension. Table 4 illustrates an example of the output generated by the prompt used for the Place dimension.

Table 4 - Example of output for the Place dimension

ID	Text	Place Aspects
12184	The hotel is situated in a very nice location near the beach with fabulous views and shaded gardens. The staff are willing and helpful. The pool is quite small and very deep. The lifts are not totally reliable. The food was not of a good standard in either quality or variety. The soft drinks that were served to Saga guests in the bar were of poor quality. These were served from a plastic bottle and if the bottle had previously been opened, the drink was flat and tasteless. We have had much better value in Spanish hotels.	[General sentiment: 1, ', Justification: The review positively mentions the hotel's location near the beach, fabulous views, and shaded gardens, which are all aspects related to the PLACE dimension. Although there are negative comments about other aspects of the hotel, such as the pool, lifts, and food quality, these do not pertain to the PLACE dimension as defined for this task. Therefore, the sentiment regarding the PLACE dimension is positive.]

Source: Own elaboration.

However, since the original dataset did not include manual human labeling, a crucial step was added to this research to ensure an evaluation of model performance. A manual human labeling process was conducted on a random sample of 500 reviews, creating a reference dataset to evaluate which model extracted the most accurate aspects and sentiments. Human validation became essential to ensure the highest accuracy of the benchmarking methodology. This approach is crucial to ensuring the integrity of the findings and demonstrates a commitment to the accuracy of the analysis. Combining human analysis with the evaluation of ChatGPT's performance makes it possible to ensure the quality of the data and the conclusions drawn.

A similarity measure, an extension of the traditional Jaccard and Dice measures (Han et al., 2011), was adopted to analyze the similarity between the three approaches. While these measures are valuable for comparing binary data, they are less effective when dealing with variables encompassing a more comprehensive range of values or non-binary representations. For that reason, we present a new measure which we call Similarity Coefficient (SC). This measure considers all dimensions as elements instead of traditional approaches that focus on individual elements.

$$Similarity\ Coefficient = \frac{Correct\ Matches}{Total\ number\ of\ Dimensions}$$

While Jaccard and Dice assess the presence or absence of specific elements in each data set, SC is intended for situations where the aim is to make objective comparisons of the number of elements shared between data sets. In this context, SC establishes a similarity scale between 0 and 1, where the minimum value (0) indicates the total absence of elements in common between the sets, while the maximum value (1) represents the presence of all elements in common. In this way, SC provides a clear and objective metric for assessing the similarity between sets, enabling a precise and quantitative analysis of the overlap between their dimensions. Its interpretation as a percentage adds a layer of flexibility, adapting to the specific objectives of each analysis.

Table 5 shows an example to illustrate the Similarity Coefficient (SC) calculation

Table 5 - Comparison values between ChatGPT and Human approaches (for a random review)

Dimension	ChatGPT	Human
Staff	-1	-1
Price	Blank	Blank
Place	-1	-1
Ambience	-1	-1
Experiences	-1	-1
Services	0	1

Source: Own elaboration.

$$SC = \frac{5}{6} \times 100 = 83$$

For the dimensions of Staff, Place, Ambience, and Experiences, both approaches rate them as -1, and for Price, both leave it blank. The only difference between the approaches is in the Services dimension, where ChatGPT rates it as 0, and the Human approach rates it as 1. It is important to note that blanks are considered valid values and are included in the calculation. With five matching dimensions out of six, the SC value of 83% indicates a high level of similarity between the two approaches across the evaluated dimensions.

4. Results

The results of comparing three distinct approaches – analysis conducted in the Rassal et al. (2023) study (Previous), human analysis (Human), and automated analysis by the ChatGPT model (ChatGPT) – unveil significant insights into the efficacy and reliability of each method. Table 6 presents the comparison values between approaches per dimension.

Table 6 - Comparison values between approaches per dimension

Dimension	ChatGPT vs Human	Human vs Previous	ChatGPT vs Previous
Staff	95%	72%	75%
Price	74%	44%	42%
Place	86%	81%	79%
Ambience	87%	71%	81%
Experiences	93%	23%	23%
Services	91%	55%	57%

Source: Own elaboration.

To contextualize these findings, the results address the study's two guiding research questions:

RQ1: How can sentiment analysis leveraging Large Language Models (LLMs) enhance the capabilities of Aspect-Based Sentiment Analysis (ABSA)?

RQ2: How does the performance of an LLM-based approach to ABSA compare to a manual analysis approach?

The results demonstrate how the LLM-based approach enhances ABSA (RQ1) and compares with the previous approach in terms of accuracy, particularly in capturing the linguistic nuances of customer reviews. In addressing RQ2, the results demonstrate that ChatGPT achieves high similarity with the Human approach, with similarity scores consistently above 70% across all dimensions and exceeding 90% in dimensions like Staff, Experiences, and Services. These findings highlight ChatGPT's ability to approximate human sentiment classification accuracy while offering the advantages of automation and scalability. However, when comparing the Previous study with Human or ChatGPT approaches, it becomes evident that their values are remarkably similar. Across various dimensions, such as Ambience, the differences are minimal, varying by a maximum of only ten percentual points, underscoring the robustness of ChatGPT.

Despite the high level of agreement between Human and ChatGPT classifications, discrepancies emerge when compared with the Previous study. The low levels of concordance in this comparison suggest that the Previous approach could have been more effective, especially in dimensions related to Price and Experiences.

A detailed analysis was conducted using the previously selected sample of 500 reviews to provide a more prominent understanding. Table 7 presents the descriptive statistics for the Similarity Coefficient (SC) values between approaches per review.

Table 7 – Descriptive Statistics for SC Values Across Comparison Approaches

Metric	ChatGPT vs Human	Human vs Previous	ChatGPT vs Previous
Maximum	100.0%	100.0%	100.0%
Minimum	64.5%	13.9%	15.2%
Mean	95.1%	56.8%	58.7%
Standard Deviation	≈ 0.081	≈ 0.152	≈ 0.143

Source: Own elaboration.

For the ChatGPT vs. Human comparison, maximum similarity reached 100%, indicating perfect agreement, while the minimum score was 64.5%, suggesting some divergence between the Human and ChatGPT approaches. Similarly, in the comparison between the Human vs. Previous studies, the maximum similarity was 100%, with a minimum score of 13.9%. In comparing the previous study and ChatGPT classifications, maximum similarity was also 100%, with a minimum score of 15.2%. The standard deviation (SD) values in Table 7 highlight the consistency of similarity scores. ChatGPT vs. Human shows the lowest SD (≈ 0.081), indicating highly consistent performance around the mean (95.1%). In contrast, higher SDs for Human vs. Previous (≈ 0.152) and ChatGPT vs. Previous (≈ 0.143) reflect greater variability, particularly in dimensions like Price and Experiences.

5. Discussion

The analysis of sentiment classification across the three approaches—Previous (Rassal et al., 2023), Human, and ChatGPT—reveals several critical insights into the performance and interpretative capabilities of each method. Discrepancies were noted predominantly in reviews where subjective nuances or conflicting polarities were present, challenging the effectiveness of traditional methods. To elaborate on these findings, a subset of the processed reviews from the sample is presented below and analyzed in detail.

Upon examining the lowest similarity values between classifications from the Previous study and the Human or ChatGPT approach, it was observed that these discrepancies occurred within the same review. See Figure 3 for example. Table 8 shows the sentiment classification for the review by dimension for the three approaches of this example.

Figure 3 – Review 1

Review 1
Room 614 cannot fault its pool area was also good. Don't waste your money going all-inclusive. It does not start till 12, which includes cold drinks, etc. The pool shack was supposed to open at 2, but it opened whenever the staff decided! Check in what an awful experience! Arrived just after 10.30 am and didn't get in our room until 16.45! We'd been traveling since 01.30, so it's not a good advert for them to have me sleep on their sofa in reception! Considering you check out at 10 they are in no rush to give you your room. I get the impression management doesn't want to cater for all-inclusive as the food wasn't great and there was not much variety. Go self-catering as loads of bars/restaurants are cheap and good quality around the hotel. The hotel charged €6 for a Coke and an orange Fanta at the pool shack! Ridiculous never again!

Source: Own elaboration.

Table 8 - Review sentiment classification

Dimension	Previous	ChatGPT	Human
Staff	1	-1	-1
Price	0	-1	-1
Place	0	Blank	0
Ambience	0	0	1
Experiences	0	-1	-1
Services	0	-1	-1

Source: Own elaboration.

Analyzing the review on the Staff dimension reveals critical issues with check-in and the pool shack's inconsistent opening times, suggesting a lack of reliability and responsiveness from the staff. The ChatGPT approach correctly identified the negative aspects associated with the staff, classifying the dimension as negative. On the other hand, the Previous approach, using the fuzzy approach, attributed a positive feeling, possibly due to the word staff in the review. However, this classification needs to be revised since the comments clearly show a negative experience with the staff.

When considering the Price dimension, dissatisfaction with the hotel's prices is evident (hotel charged €6 for a coke and an orange fanta at pool shack!). The previous approach assigned a neutral value to this dimension since it is not explicitly mentioned. In contrast, ChatGPT correctly identified the dissatisfaction expressed in relation to prices, assigning a negative sentiment to the price dimension since it was based on the context provided in the review. The last sentence reinforces dissatisfaction with the hotel's prices.

In the case of the Place dimension, it was considered neutral by the Human and Previous approach, while ChatGPT classified it as Blank, indicating the absence of aspects for this dimension. The difference in rating can be justified by the lack of specific mentions of aspects related to the Place dimension, as justified by ChatGPT. However, this assessment can be contested since the expression Room 614 can not fault it indirectly indicates satisfaction with the quality of the room.

As for the Ambience dimension, the divergent classification between approaches suggests that the interpretation of feelings may vary depending on the emphasis given to different aspects of the review. While the Human approach assigned a positive sentiment, ChatGPT considered the review neutral due to the mixed comments about the overall experience at the hotel. These discrepancies highlight the importance of considering subjectivity and context.

For the Services dimension, the review indicates dissatisfaction with the all-inclusive service due to the lack of cold drinks and poor quality and variety of food. In this situation, where the review is not objective, the Previous approach failed to classify adequately, assigning a neutral sentiment, unlike ChatGPT, which agrees with the Human approach, with both negative evaluations.

The same occurs in the Experiences dimension. The review reflects the customer's negative experience and lack of desire to return, but the previous study incorrectly gave a neutral sentiment. The fact is that the Experience dimension in the Previous approach is only 23% in line with Human and ChatGPT approaches. This discrepancy is related to the methodology employed by the previous approach, which assigned an average degree of pertinence when considering both positive and negative points. However, the review's predominance of negative aspects results in an overall negative experience, where unfavorable aspects outweigh positive ones. ChatGPT correctly identified this trend, emphasizing the importance of considering the whole context.

The difficulty in correctly categorizing the Experience dimension also arises when a positive experience is in question. Considering the review in Figure 4, the result for approach was: Blank, 1, and 1, respectively for Previous, ChatGPT, and Human.

Figure 4 – Review 2

Review 2
It may not have the trappings of a 5-star hotel, but it offers fantastic value for your money with extremely comfortable rooms that are well equipped if you want to just stay in and enjoy the view from the balcony or a great base to explore both the old and new town. We stayed here for a week, which must have been good already booked two weeks for next February, and must say we had no problems with either the staff or service, ate a few nights in the Hotel restaurant and the food was good, my only complaint was the wifi which was abysmal in the room but functional if you were in reception. If that's my only complaint, it must have been good. I would heartily recommend the hotel (…)

Source: Own elaboration.

The review highlights several positive aspects of the hotel experience, such as comfortable rooms and a favorable location for exploring the city, and praising the service's quality. Although he mentioned a minor complaint about the Wi-Fi, the general feeling remained positive. In the ChatGPT analysis, the review was interpreted as positive regarding the Experiences dimension, similar to the Human approach. Although the problem with the Wi-Fi was mentioned, it did not affect the overall positive feeling expressed in the review. It is interesting to note that the above approach indicates the absence of mention of the Experiences dimension. This difficulty in identifying and assigning a score to the Experiences dimension emphasizes the advantage of using LLMs, which can process a large amount of contextual information and linguistic nuances, as well as understand the whole context.

When analyzing other reviews, the previous study's difficulty in adequately interpreting the context of sentences becomes clear. See Figure 5 for example. The result for approach was -1, 1 and 1, respectively for Previous, ChatGPT, and Human.

Figure 5 – Review 3

Review 3
(…) it's a beautiful location, and the hotel, food and staff are also great. unfortunately, our room is very disappointing. It's very cold and damp, both single beds, and now my clothes are damp. There is no sun at all in our balcony or room ever it's very uncomfortable to come into a really cold room and sleep in separate cold beds (…)

Source: Own elaboration.

In the previous review, where it is mentioned that the staff is great, the Previous study made a mistake by wrongly attributing a negative feeling to the staff dimension. This discrepancy highlights a need to understand the comment's true meaning, omitting the sentence's full context. By failing to recognize the expression of a positive sentiment towards the staff, the previous analysis reinforced its limitation in accurately interpreting the sentiments expressed by customers, failing to deal with the absence of punctuation in the sentence. This failure to recognize the duality of the feelings expressed in the review emphasizes the need for more sophisticated methods of textual analysis capable of adequately understanding and interpreting the complexity of human expressions, such as LLMs, underscoring the significance of the present study.

Figure 6 provides an additional example highlighting ChatGPT's distinct ability to interpret a sentence's context. The sentiment analysis of the review by the Price dimension was -1, Blank and Blank, respectively for Previous, ChatGPT and Human approach.

Figure 6 - Review 4

Review 4
(…) There were 2 snack bars, but only one was open in the main building, which made it a bit difficult to get a snack if you were at the swimming pool, which was on the other side, but it was not a problem! Snack bar was opened 24/7 which was great! (…) Overall we were very happy, and we will visit the hotel again! Near the hotel there are 3 lovely beaches, just 5 min walk away!! Really, really stunning! You can also take a boat trip for approximately 18 euros and it lasts 1 hour, it's really, really worth it! There are many activities you can do every day with the activity team (…)

Source: Own elaboration.

In this review, the customer discusses an activity outside the hotel - a boat trip - and expresses satisfaction with the perceived value of this activity. In the context of the Price dimension, which focuses on aspects related to costs within the hotel, both ChatGPT and the Human approach correctly identify that this mention is not directly associated with the hotel establishment in question. Consequently, a blank rating is given to the price dimension. On the other hand, the previous study failed to make this distinction, leading to a negative sentiment rating. Due to their high capacity to deal with subjectivity and ambiguity more effectively, learning-based language models such as LLMs often outperform humans in interpreting complex texts.

Considering the review in Figure 7, the result for the approach was: 1, 1, and Blank, respectively, for Previous, ChatGPT, and Human.

Figure 7 - Review 5

Review 5
The modern suite met most of our standards: spacious living, nice kitchen, with all necessary equipment, two separate bedrooms with bathrooms each, and a nice sea view too (…). The apartment was cleaned every day and service was friendly. We would have liked, though, to have Wifi not just at the reception but inside the apartment as well. And we didn't manage to get the heating on, while the temperature outside dropped to 14 degrees in November. Hopefully, it will work next year (…).

Source: Own elaboration.

While the Human approach classified the Staff dimension as blank, indicating the absence of sentiment, ChatGPT interpreted the mention of service as positive, highlighting the friendliness of the service. In the justification provided by ChatGPT, the mention of friendly service is interpreted as an indicator of positive sentiment towards the hotel staff, highlighting a subtlety in the language that may have been overlooked in the Human analysis.

This discrepancy between the interpretations accents one of the nuances of sentiment analysis, where subjectivity and context play a key role. While the Human approach may be more rigid in its classification, ChatGPT demonstrates a greater sensitivity to detail and the interrelationship between the different aspects, holding a significant promise for enhancing ABSA.

Unlike previous studies, such as those by Rathan et al. (2018) and Pham and Le (2018), ChatGPT considers both the words used and the broader context and implicit connections between different elements. This capability accentuates the importance of a flexible and sensitive approach, in which tools such as ChatGPT can offer valuable insights that could easily be overlooked in more rigid Human analysis.

Figure 8 is another example that highlights this capability. The results for the Ambience dimension were 1, 0, and 1, respectively, for the Previous, ChatGPT, and Human approaches.

Figure 8 - Review 6

Review 6
(…) We got a studio apartment which was adequate for our needs, basic but clean with a balcony. The maids clean the room regularly and change the towels and sheets regularly. My only disappointment was that I was unable to swim in the pool every day. The sun loungers get snapped up quickly, and the pool gets really crowded with people playing with bats and balls. My tip would be to put your towel down on the way to breakfast! (…) This complex is ideal for families who want an all-inclusive holiday as there is also plenty for the kids to do (…)

Source: Own elaboration.

ChatGPT justified its rating in this review by considering both positive and negative aspects. On the one hand, the flat is described as adequate for our needs, basic but clean with a balcony, indicating a positive experience with the accommodation. However, the crowdedness of the pool area (The sun loungers get snapped up quickly) is mentioned as a negative aspect, which may influence guests' overall experience of the environment. While the Human approach may have been influenced predominantly by the positive aspects, ChatGPT recognized the presence of negative aspects that affect the guest's overall perception of the environment. This situation illustrates how ChatGPT can be more sensitive to specific nuances and provide a more complete analysis in certain contexts.

Conversely, as shown in Figure 9, there are situations that require objectivity. The results for the three approaches were -1 for the Previous approach, 0 for ChatGPT, and 1 for the Human approach.

Figure 9 - Review 7

Review 7
(…) it's a beautiful location, and the hotel, food, and staff are also great. unfortunately, our room is very disappointing. It's very cold and damp, both single beds, and now my clothes are damp. There is no sun at all in our balcony or room ever it's very uncomfortable to come into a really cold room and sleep in separate cold beds, it was meant to be a holiday for rest and togetherness, but I can't wait to get home.

Source: Own elaboration.

For example, in Review 7, it is possible to see a situation in which the complexity and subtlety of human language can influence ChatGPT's interpretation. While the Human approach attributed a positive sentiment to the services dimension due to the praise for the food, ChatGPT, when considering the broader context of the review, attributed a neutral sentiment, considering the room conditions for the classification of the Services dimension. However, it is important to note that aspects related to the room have already been discussed in the Place dimension, which specifically addresses the physical characteristics and location of the hotel.

In conclusion, although LLMs positively addressed the two research questions (RQ1 and RQ2), the nuanced discrepancies observed between the Human and ChatGPT approaches highlight the importance of meticulous and complementary analysis, particularly in contexts where the subtleties of human language and contextual understanding play a crucial role in interpretation.

5.1 *Theoretical Implications*

This study underscores the advantages of Large Language Models (LLMs) like ChatGPT in capturing nuanced sentiments and resolving ambiguities, which traditional methods such as fuzzy logic often struggle to address. For instance, discrepancies in classifications related to the Staff and Price dimensions highlight the importance of holistic sentiment understanding. ChatGPT's ability to interpret implicit relationships and contextual subtleties could mark a significant step forward in Aspect-Based Sentiment Analysis (ABSA). By addressing the limitations of earlier methods, ChatGPT facilitates a more detailed understanding of customer feedback, enabling researchers to examine multiple sentiments expressed within a single review. This capability is particularly important in contexts where subjectivity and ambiguity prevail, as it bridges the gap between computational efficiency and human-like interpretative depth. These findings reinforce prior research by Simmering & Huoviala (2023) and Macháliková (2023), which emphasized the effectiveness of LLMs in nuanced sentiment classification and provide a base for future explorations, encouraging the adoption of LLM-based solutions across diverse research domains.

5.2 Practical Implications

From an application perspective, ChatGPT proves to be a scalable and efficient tool for sentiment analysis, offering performance comparable to manual human evaluations while reducing resource intensity. Industries such as hospitality and tourism, which depend heavily on customer feedback for service improvements, can leverage LLMs to analyze sentiments more effectively and efficiently. This capability facilitates the extraction of detailed insights and the automation of feedback analysis, enabling quicker and more informed decision-making. By incorporating ChatGPT into sentiment analysis workflows, organizations can achieve better responsiveness and alignment with customer needs. The findings also suggest that ChatGPT's integration into customer service platforms could significantly enhance sentiment classification accuracy, leading to improved customer satisfaction and loyalty.

Furthermore, ChatGPT's ability to identify nuanced feedback opens opportunities for personalized recommendations and targeted marketing strategies. As highlighted in the literature (e.g., Liu et al., 2023), these insights align with the growing evidence supporting the integration of LLM-based solutions into practical applications. The scalability and adaptability of such models could provide a cost-effective and high-performance alternative to traditional sentiment analysis methods, making them particularly valuable for large-scale datasets.

6. Conclusions

In today's business environment, understanding customer feedback in a detailed way is essential. The process becomes more challenging when a single sentence addresses multiple aspects and includes opinions with opposing polarities. This is where ABSA comes in. However, due to the complexity of human language and the time required, traditional sentiment analysis methods became unreliable. They often failed to capture customer sentiment accurately, especially when dealing with subjective and ambiguous language. This study uses GPT-4o, an LLM, to improve ABSA by taking advantage of ChatGPT's capabilities. Two approaches were employed to benchmark the novel ChatGPT approach—a Previous approach (based on a dictionary approach) and a Human approach—to determine how LLMs can be used to enhance ABSA.

Comparing the GPT-4o approach with the two other approaches revealed a high similarity between the Human and ChatGPT approaches, consistently above 70% across all dimensions, particularly in the Staff, Experiences, and Service dimensions, exceeding 90%. These findings highlight ChatGPT's capacity to understand context and implicit relationships in customer reviews. The model can process large volumes of data, providing a more refined sentiment classification than traditional methods.

This study offers a more flexible and sensitive approach to sentiment analysis, marking significant achievements by overcoming the limitations of previous techniques, and represents an advancement in the field through its novel application of LLMs for ABSA in customer reviews. Building on these foundations, the study validates the claims of Simmering and Huoviala (2023), who highlighted the superiority of LLMs over traditional methods, and corroborates Machálikova's (2023) conclusions about the effectiveness of ChatGPT, contributing to the growing evidence supporting LLMs in ABSA. Additionally, this study establishes a methodology for how other researchers can use LLMs for ABSA studies.

These findings contribute to academic research and hold implications for organizations, especially tourism-related organizations, that seek efficient and scalable solutions for customer feedback analysis and market research. With this solution, organizations can conduct a more precise and rapid sentiment analysis using ChatGPT to extract insights from vast amounts of data, allowing them to understand areas of improvement. Moreover, the automated review analysis and tailored recommendation generation lead to enhanced customer loyalty.

Despite the advancements, it is crucial to acknowledge some limitations. While ChatGPT demonstrates impressive capabilities in simulating human-like sentiment analysis, its interpretations may be influenced by the quality and specificity of the prompt provided. Furthermore, the study found instances where executing the code produced different results with the same command. This variability highlights the complexity of LLM models, which rely on probabilistic processes and may yield divergent results under certain conditions. Using a sample of 500 reviews and the only use of ChatGPT through GPT-4o also could limit the representativeness of the training data, affecting the model's quality.

Following these limitations, future efforts may focus on increasing the sample size and extending it to other contexts, as well as using other versions of GPT or other LLMs such as Gemini, Claude 3.5, or even open-source options such as Mistral and LLama 3, with the advantage of being open source and cost-free. Additionally, providing more training examples and simplified prompts could be crucial for enhancing the quality of the results. Finally, establishing benchmarks with other LLMs can guide researchers in selecting the most suitable model for specific tasks.

Credit author statement

All authors have contributed equally. All authors have read and agreed to the published version of the manuscript.

Declaration of competing interest: None

Acknowledgements

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

References

Aguilar-Moreno, J. A., Palos-Sanchez, P. R., & Pozo-Barajas, R. D. (2024). Sentiment analysis to support business decision-making: A bibliometric study. *AIMS Mathematics*, 9(2), 4337–4375. https://doi.org/10.3934/math.2024215

Al-Smadi, M., Talafha, B., Al-Ayyoub, M., & Jararweh, Y. (2019). Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *International Journal of Machine Learning and Cybernetics*, 10(8), 2163–2175. https://doi.org/10.1007/s13042-018-0799-4

Amar , J., I., Muna Almaududi Ausat, A., Sumarna, A., Studi Magister Manajemen, P., & Tinggi Ilmu Ekonomi YPUP Makassar, S. (2023). Application of ChatGPT in Business Management and Strategic Decision Making. *Jurnal Minfo Polgan*, 12(2). https://doi.org/10.33395/jmp.v12i2.12956

API Reference - OpenAI API. (n.d.). Retrieved April 3, 2024, available from https://platform.openai.com/docs/api-reference/authentication

Areed, S., Alqaryouti, O., Siyam, B., & Shaalan, K. (2020). Aspect-Based Sentiment Analysis for Arabic Government Reviews. *In Studies in Computational Intelligence*, 874, 143–162. https://doi.org/10.1007/978-3-030-34614-0_8

Azam, M., Raiaan, K., Saddam, M., Mukta, H., Fatema, K., Fahad, N. M., Sakib, S., Mim, J., Ahmad, J., Ali, M. E., & Azam, S. (2023). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *techRxiv preprint techrxiv:10.36227.* https://doi.org/10.36227/techrxiv.24171183

Baker, M. R., & Utku, A. (2023). Unraveling user perceptions and biases: A comparative study of ML and DL models for exploring twitter sentiments towards ChatGPT. *Journal of Engineering Research*. https://doi.org/10.1016/j.jer.2023.11.023

Banjar, A., Ahmed, Z., Daud, A., Abbasi, R. A., & Dawood, H. (2021). Aspect-Based Sentiment Analysis for Polarity Estimation of Customer Reviews on Twitter. *Computers, Materials and Continua*, 67(2), 2203–2225. https://doi.org/10.32604/cmc.2021.014226

Carvalho, F., Ramos, R. F., & Fortes, N. (2024). Customer satisfaction in mountain hotels within UNESCO Global Geoparks: an empirical study based on sentiment analysis of online consumer reviews. *Tourism & Management Studies, 20(1)*, 35-47. https://doi.org/10.18089/tms.20240103

Chifu, A.-G., & Fournier, S. (2023). Sentiment Difficulty in Aspect-Based Sentiment Analysis. *Mathematics*, 11(22), 4647. https://doi.org/10.3390/math11224647

Dasgupta, D., Venugopal, D., & Gupta, K. D. (2023, February). A review of generative AI from historical perspectives. *techRxiv preprint techrxiv:10.36227.* https://doi.org/10.36227/techrxiv.22097942.v1

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science*, 87, 44–49. https://doi.org/10.1016/j.procs.2016.05.124

George, A. S., George, A. S. H., & Martin, A. S. G. (2023). A Review of ChatGPT AI's Impact on Several Business Sectors. *Partners Universal International Innovation Journal*, 1(1), 9—22, https://doi.org/10.5281/zenodo.7644359

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques (3rd ed.).* Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-Based Sentiment Analysis using BERT. *In Proceedings of the 22nd Nordic Conference on Computational Linguistics,* 187–196. https://aclanthology.org/W19-6120.pdf

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. Department of Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7053. @cs.uic.eduHussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences, 30(4)*, 330–338. https://doi.org/10.1016/j.jksues.2016.04.002

Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, 30(4), 330–338. https://doi.org/10.1016/j.jksues.2016.04.002

Ismet, H. T., Mustaqim, T., & Purwitasari, D. (2022). Aspect Based Sentiment Analysis of Product Review Using Memory Network. *Scientific Journal of Informatics*, 9(1), 73–83. https://doi.org/10.15294/sji.v9i1.34094

Khalaf, I., Al-Tameemi, S., Feizi-Derakhshi, M.-R., Pashazadeh, S., & Asadpour, M. (2022). A Comprehensive Review Of Visual-Textual Sentiment Analysis From Social Media Networks. *preprint arXiv:2207.02160.* https://doi.org/10.48550/arXiv.2207.02160

Kheiri, K., & Karimi, H. (2023). Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234.* https://doi.org/10.48550/arXiv.2307.10234

Kontonatsios, G., Clive, J., Harrison, G., Metcalfe, T., Sliwiak, P., Tahir, H., & Ghose, A. (2023). FABSA: An aspect-based sentiment analysis dataset of user reviews. *Neurocomputing*, 562, 126867. https://doi.org/10.1016/j.neucom.2023.126867

Korkmaz, A., Aktürk, C., & Talan, T. (2023). Analyzing the User's Sentiments of ChatGPT Using Twitter Data. *Iraqi Journal for Computer Science and Mathematics*, 202–214. https://doi.org/10.52866/ijcsm.2023.02.02.018

Kumar, A., & Sebastian, T. M. (2012). Sentiment analysis on Twitter. *International Journal of Computer Science Issues*, 9(4), 372. Retrieved from http://www.IJCSI.org76Li, H., Yu, B. X. B., Li, G., & Gao, H. (2023). Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews. *Tourism Management*, 96. https://doi.org/10.1016/j.tourman.2022.104707

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167. https://doi.org/10.1007/978-3-031-02145-9

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

Machálíková, K. (2023). Utilizing ChatGPT for Sentiment Analysis Title Utilizing ChatGPT for Sentiment Analysis. Retrieved from https://urn.fi/URN:NBN:fi:amk-2023121537710

Mojica, M., Palos-Sanchez, P.R. & Cabanas, E. (2024), "Is there innovation management of emotions or just the commodification of happiness? A sentiment analysis of happiness apps", *European Journal of Innovation Management*, * Advance online publication. https://doi.org/10.1108/EJIM-11-2023-0963

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., & Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers*. IEEE Access*, 8, 131662–131682. https://doi.org/10.1109/ACCESS.2020.3009626

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435.* https://arxiv.org/abs/2307.06435

Nkhata, G. (2022). Movie Reviews Sentiment Analysis Using BERT. Graduate Theses and Dissertations Retrieved from https://scholarworks.uark.edu/etd/4768

Oliveira, A. S., Renda, A. I., Correia, M. B., & Antonio, N. (2022). Hotel customer segmentation and sentiment analysis through online reviews: An analysis of selected European markets. *Tourism & Management Studies*, 18(1), 29-40. https://doi.org/10.18089/tms.2022.180103

OpenAI API. (n.d.). Authentication. Retrieved from https://platform.openai.com/docs/api-reference/authentication

Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (ACL-04), 271–278. https://doi.org/10.3115/1218955.1218990

Perea-Khalifi, D., Irimia-Diéguez, A. I., & Palos-Sánchez, P. (2024). Exploring the determinants of the user experience in P2P payment systems in Spain: A text mining approach. *Financial Innovation*, 10(1), 2. http://dx.doi.org/10.1108/EJIM-11-2023-0963

Pham, D. H., & Le, A. C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data & Knowledge Engineering*, 114, 26–39. https://doi.org/10.1016/J.DATAK.2017.06.001

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis*. International Workshop on Semantic Evaluation*. https://doi.org/10.3115/v1/S14-2004

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. *International Workshop on Semantic Evaluation*. http://doi.org/10.18653/v1/S15-2082

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I., Bel, N., & Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *International Workshop on Semantic Evaluation*. https://doi.org/10.18653/v1/S16-1002

Raj, R., Singh, A., Kumar, V., & Verma, P. (2023). Analyzing the potential benefits and use cases of ChatGPT as a tool for improving the efficiency and effectiveness of business operations. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(3), 100140. https://doi.org/10.1016/j.tbench.2023.100140

Rassal, C., Correia, A., & Serra, F. (2023). Understanding Online Reviews in All-Inclusive Hotels Servicescape: A Fuzzy Set Approach. *Journal of Quality Assurance in Hospitality & Tourism*, 25(6), 1607–1634. https://doi.org/10.1080/1528008X.2023.2167761

Rathan, M., Hulipalled, V. R., Venugopal, K. R., & Patnaik, L. M. (2018). Consumer insight mining: Aspect-based Twitter opinion mining of mobile phone reviews. Applied Soft Computing, 68, 765-773. https://doi.org/10.1016/j.asoc.2017.07.056

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, 6(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311, 18–38. https://doi.org/10.1016/j.ins.2015.03.040

Simmering, P. F., & Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. *arXiv preprint arXiv:2310.18025*. https://doi.org/10.48550/arXiv.2310.18025

Sudirjo, F., Diantoro, K., Al-Gasawneh, J. A., Khootimah Azzaakiyyah, H., & Almaududi Ausat, A. M. (2023). Application of ChatGPT in Improving Customer Sentiment Analysis for Businesses*. Jurnal Teknologi Dan Sistem Informasi Bisnis*, 5(3), 283–288. https://doi.org/10.47233/jteksis.v5i3.871

Tan, T. F., Thirunavukarasu, A. J., Campbell, J. P., Keane, P. A., Pasquale, L. R., Abramoff, M. D., Kalpathy-Cramer, J., Lum, F., Kim, J. E., Baxter, S. L., & Ting, D. S. W. (2023). Generative Artificial Intelligence Through ChatGPT and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges. *Ophthalmology Science*, 3(4). https://doi.org/10.1016/j.xops.2023.100394

Thet, T. T., Na, J. C., & Khoo, C. S. G. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6), 823–848. https://doi.org/10.1177/0165551510388123

Truşcă, M. M., & Frasincar, F. (2023). Survey on Aspect Detection for Aspect-Based Sentiment Analysis. *Artificial Intelligence Review*, 56, 3797–3846. https://doi.org/10.1007/s10462-022-10252-y

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wang, B., & Liu, M. (2022). Deep learning for aspect-based sentiment analysis. *PeerJ Computer Science*, 8, 1—37. https://doi.org/10.7717/peerj-cs.1044

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Yñiguez-Ovando, R., Buitrago-Esquinas, E. M., Puig-Cabrera, M., Santos, M. C., & Santos, J. A. C. (2024). Artificial Intelligence and Sustainable Tourism Planning: A Hetero-Intelligence Methodology Proposal. *Tourism & Management Studies*, 20(SI), 45-59. https://doi.org/10.18089/tms.2024SI04

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*. https://doi.org/10.48550/arXiv.2305.15005