IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 June 2024, accepted 10 July 2024, date of publication 16 July 2024, date of current version 26 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3429290

RESEARCH ARTICLE

Harnessing the Power of LLMs for Service Quality Assessment From User-Generated Content

TAHA FALATOURI[1,2,3], **DENISA HRUŠECKÁ**[1], **AND THOMAS FISCHER**[2,3]
[1]Faculty of Management and Economics, Tomas Bata University in Zlín, 760 01 Zlín, Czech Republic
[2]Department for Logistics, University of Applied Sciences Upper Austria, 4400 Steyr, Austria
[3]Josef Ressel-Centre for Predictive Value Network Intelligence, 4400 Steyr, Austria

Corresponding author: Taha Falatouri (taha.falatouri@fh-steyr.at)

This work was supported by the Christian Doppler Research Association as part of the Josef Ressel Centre for Predictive Value Network Intelligence (JRC PREVAIL).

ABSTRACT Adopting Large Language Models (LLMs) creates opportunities for organizations to increase efficiency, particularly in sentiment analysis and information extraction tasks. This study explores the efficiency of LLMs in real-world applications, focusing on sentiment analysis and service quality dimension extraction from user-generated content (UGC). For this purpose, we compare the performance of two LLMs (ChatGPT 3.5 and Claude 3) and three traditional NLP methods using two datasets of customer reviews (one in English and one in Persian). The results indicate that LLMs can achieve notable accuracy in information extraction (76% accuracy for ChatGPT and 68% for Claude 3) and sentiment analysis (substantial agreement with human raters for ChatGPT and moderate agreement with human raters for Claude 3), demonstrating an improvement compared to other AI models. However, challenges persist, including discrepancies between model predictions and human judgments and limitations in extracting specific dimensions from unstructured text. Whereas LLMs can streamline the SQ assessment process, human supervision remains essential to ensure reliability.

INDEX TERMS ChatGPT, cloud 3, large language models (LLMs), natural language processing (NLP), sentiment analysis, service quality assessment.

I. INTRODUCTION

In the era of online businesses, there is growing recognition of service quality (SQ) shaping customer satisfaction [1]. User-generated content (UGC) can be a crucial resource for companies aiming to gain deeper insights into their SQ [2]. UGC has been recognized as a critical component of online word-of-mouth (WOM), enabling customer feedback collection before, during, and after service interactions [3]. The trend of using pre-trained language representations in natural language processing (NLP) systems has received colossal attention recently [4]. Companies are also trying to benefit from Large language models (LLMs) in their daily business [5]. Despite the importance of SQ and the widespread use of LLMs, the effectiveness of these methods in this area has not been investigated [1].

Most of the literature related to assessing SQ has employed pre-defined questionnaires, limiting surveys to common dimensions. In contrast, UGC is often developed under less controlled circumstances and frequently after considerable contemplation, resulting in broader outcomes. In addition, UGC minimizes potential biases as the contributors are unaware of their participant role [1], [6]. Nevertheless, although UGC presents appealing opportunities, its use in assessing service quality remains limited [7]. A potential explanation for this circumstance is the need to deal with a large quantity of loosely structured data when trying to derive insights from UGC.

Hence, data analytic advancements are needed that enable an analysis of this kind of data, allowing domain experts (e.g., sales experts or marketing managers) who may not

The associate editor coordinating the review of this manuscript and approving it for publication was Maria Chiara Caschera.

have the technical expertise to implement the complex procedures themselves. In the present Artificial Intelligence (AI) era, advanced language models such as OpenAI's ChatGPT present a promising avenue for text classification and sentiment analysis [8]. Having undergone rigorous training on diverse datasets, these advanced models have demonstrated exceptional proficiency in understanding and generating text that closely mirrors human language [9], [10]. Before LLMs, the conventional approach for text mining tasks involved data preparation and computationally intensive fine-tuning. Generative LLMs eliminate the need for fine-tuning by using specific prompts, smoothing the process [11]. As this category of applications has now also widely arrived in the daily business of large companies worldwide (e.g., with 92% of Fortune 500 members incorporating ChatGPT into their daily tasks, 12), the question arises whether LLMs can be a way to deal with UGC for company representatives to get a better idea of the service quality perceptions of their customers. Even though they demonstrate poorer performance when tackling complex tasks [13].

Related to these research interests and the need for further investigation into the role of UGC and LLMs in the assessment of service quality, we aim to answer the following research questions as part of the study presented in this article:

RQ: *How can an LLM contribute to the assessment of SQ from UGC*?

II. RESEARCH BACKGROUND

A. SERVICE QUALITY (SQ) ASSESSMENT

Over the past decades, the growth of e-commerce and increased competitiveness in the retail market have prompted market participants to operate more efficiently than before. Retailers are now compelled to emphasize customer satisfaction more than in the past. In service industries, customer satisfaction and service quality have emerged as primary focal points over the last two decades [14]. It is widely recognized that service quality is essential for achieving customer satisfaction. This requires ongoing efforts to attract, satisfy, and retain customers in profit and non-profit organizations [15]. These efforts rely on data that enables managers to track current customer perceptions of service quality. For this purpose, various measurement models and related self-report instruments have been developed.

One of the initial methods to assess Service Quality is the Gap model formulated by Parasuraman et al. [16] called SERVQUAL [16]. Following the conceptualization and measurement of SERVQUAL, Cronin and Taylor [17] introduced SERVPERF. This approach sees service quality as a consumer attitude, advocating for a performance-only measurement as a more effective way to assess SQ [17]. Dabholkar et al. [18] proposed an alternative SQ model for technology-based self-service options, incorporating dimensions such as physical aspects, reliability, personal interaction, and problem-solving and policy measures (RSQS). They highlighted that RSQS emphasizes customers' perceptions of a service rather than their expectations [18]. CALSUPER incorporates service level and product quality into the SQ framework; this model is particularly applicable in supermarket settings [19]. Finally, the E-S-Qual and E-RecS-QUAL have been introduced to assess the SQ in online stores, with adding dimensions focusing on online availability and fulfillment [20], [21]. More recent papers attempt to evaluate the importance and usability of these dimensions in different areas, as shown by Mamakou et al. [22], who explored the interaction between electronic service quality and user experience; key dimensions of service quality in mobile shopping (m-shopping) have been evaluated by Zhang et al. [23], the impact of particular dimensions of logistics services on product satisfaction within the e-commerce sector was investigated by Rashid and Rasheed [24].

It is apparent that specific models and measurement instruments are available for a wide variety of different sub-domains of service quality, mainly depending on the context that one is interested in (e.g., supermarket setting compared to online retail). These potential differences in the dimensionality of service quality and the effort required to gather data based on self-reports continuously raise the need for more flexible and less obtrusive ways to gather information on customers' SQ perceptions. A viable approach in this context is the use of less structured sources of data that are directly created by customers, such as customer reviews of products, commonly referred to as UGC.

UGC refers to content created by everyday individuals who willingly provide data, facts, or media, which is then shared with others in a beneficial or enjoyable manner, typically on the internet. Examples include restaurant reviews, collaborative websites, and videos [25]. These contents can differ widely in terms of their characteristics; for example, they may involve different languages, use different jargon, or include more subtle aspects of language such as sarcasm. Further, these forms of content can often be more than pure text (e.g., emojis and gifs can be included). These variations are further complicated by platform-specific options and data entry limitations (e.g., allowed text patterns and lengths). Hence, data analytic approaches are needed to make sense of this type of data. For example, content analysis [26], [27] and topic modeling [28], [29] are common methods in this area. More recently, machine learning models have been proposed to enhance the speed and quality of UGC [30]. As a further step in making such forms of analysis more approachable for domain experts, LLMs are investigated as another data analytic avenue as part of the research presented in this article.

B. LARGE LANGUAGE MODELS (LLMS)

Traditional natural language processing (NLP) models have been widely used and studied for decades and are still relevant

in some contexts (see, for example, 8 for an overview). However, they often require manual feature engineering and may struggle to capture complex linguistic patterns compared to deep learning models like LLMs. Traditional models typically rely on features, rule-based systems [31], or statistical techniques such as N-gram models that predict the probability of a word given the previous (n-1) words and are mainly used for text correcting. Another example is rule-based models, which use linguistic knowledge and heuristic rules to process and analyze text, such as through regular expressions [32]. Rule-based models are characterized by their simplicity, relying on manually designed rules and sentence patterns. These models are limited by the predetermined knowledge base and rules provided to them. Statistical methods are the most widely used language models, like TF-IDF (Term Frequency-Inverse Document Frequency) [33] and vector space techniques used to represent language elements as vectors positioned within specific vector spaces. These models are commonly used for document retrieval, ranking, and search engine indexing tasks.

Large Language Models are advanced AI models designed to understand and generate human-like text [34]. LLMs like GPT (Generative Pre-trained Transformer) and pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) are trained on vast amounts of text data. They can perform various NLP tasks such as text generation, language translation, sentiment analysis, question answering, and more. They have significantly advanced the state-of-the-art in NLP and are widely used in various applications and research fields. By Utilizing the self-attention mechanism[1] LLMs enhance the scalability and efficiency of textual analysis [35]. Some examples of current LLMs are BERT [34], RoBERTa [36], or OpenAI's solution series, which further expanded the abilities of LLMs by incorporating the self-attention mechanism [37], [38], [39]. It has been argued that LLMs amplify the capacity of systems to analyze and manipulate text [40], which has also been demonstrated for various applications that deal with textual data, such as machine translation, language translation, text summarization, and natural language processing [37].

C. POTENTIAL OF LLMS FOR SQ ASSESSMENT

How SQ assessment can be supported by LLMs now depends on the tasks that are regarded as part of SQ assessment and their respective requirements. SQ assessment generally identifies the gap between customer expectations and perceived service [16]. To identify and measure this gap, we define two steps: (1) Determine if service quality is being addressed and, ideally, specify which aspect of service quality; (2) Assess how well the service quality (and its aspects) meets customer expectations. Data analytic approaches using LLMs can be implemented for both. As an additional condition, we want to use already existing, user-generated data as the input for these steps, as they can be an excellent source to identify service quality-related issues (e.g., 41).

The first step, the *Extraction of SQ-related Content*, could be implemented using content analysis [27] or traditional NLP models [42]. Alternatively, LLMs are able to categorize text into predefined categories or classes [34]. This can be done by providing some examples for specific categories [43]. The potential of such an approach, also based on UGC, has, for example, been previously demonstrated by Alexander et al. [44] in the context of healthcare service feedback.

For the second step, we want to assess whether customers were satisfied with the aspects of service quality mentioned in the input data based on the results of step 1. For this purpose, we are interested in the potential of sentiment analysis to be implemented using LLMs. *Sentiment analysis* deals with assessing affective responses (e.g., positive or negative) through text analysis, which can be of great potential value in the context of the large amounts of UGC created online [45], [46]. Extracting this kind of polarity from text has also been demonstrated to be possible in principle using LLMs by previous research [34], [47]. As there is initial evidence for the usefulness of LLMs, we also want to demonstrate their practicality for this purpose. We, therefore, use UGC to more closely reflect the already available data to many companies to assess their service quality. In addition, we also use cross-lingual data to reflect the potential international origin of the customer base of many companies. LLMs are also helpful in this specific context, as they can undergo training on datasets containing multiple languages, enabling them to classify text across various linguistic contexts [48]. The particular linguistic context we are investigating here, in addition to English, is a Middle Eastern language, namely Persian, as most of the literature in the context of NLP has so far focused on Western languages, and there is limited research on their application in middle eastern languages [49], [50].

III. RESEARCH METHODOLOGY

To assess the suitability of an LLM for SQ assessment, we extracted and analyzed SQ-related comments directly from UGC. Traditionally, text mining processes for extracting and analyzing comments consist of a multi-step approach. Regardless of the specific task, the data preparation first involves a tokenization process, where the entire document is divided into words, and stop words are removed using various libraries. Additionally, stemming is employed, typically transforming verbs into their roots. Finally, there is a step involving the modification of the part of speech [51]. Whereas the new generation of text mining libraries has automated the process, these generic methods often present numerous difficulties and may inadvertently remove meaningful parts of the text. On the other hand, using pre-trained LLMs can

[1]The self-attention mechanism efficiently takes into account the representations of the positional relationships or distances among elements within a sequence (Shaw, Uszkoreit, & Vaswani, 2018).

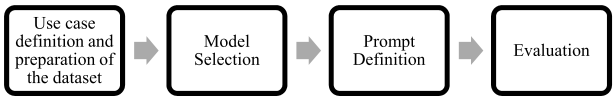

FIGURE 1. Steps in text mining – overview.

ease this process [52]. The steps leading to this objective are illustrated in Figure 1.

A. USE CASE DEFINITION AND PREPARATION OF THE DATASET

We gathered two datasets to demonstrate the suitability of LLMs for SQ assessment, including customer reviews of mobile apps, with Dataset 1 being in English and Dataset 2 being in Persian. Dataset 1 is available on Kaggle,[2] includes more than 12,000 reviews of mobile apps drawn from Google's Playstore. Dataset 2 includes more than 8,000 reviews of the mobile app of an Iranian supermarket (OKALA) and more than 1,000 reviews of the app for the customer club of the same supermarket (OK Club). The reviews were gathered from cafebazaar.ir, which is Iran's most popular website for rating Android apps (see Figure 2 For an example of the web interface). This dataset was acquired in February 2022 and covers reviews from April 2019 to February 2022. The reviews obtained had a maximum length of 300 characters (due to a length limitation of the website the reviews were gathered from).

Both datasets include ID, username, comment, star rating, and date. The comment could vary from a space character to a full text, including emojis and other graphical elements.

In the initial cleaning stage, we removed comments containing no text. Using state-of-the-art LLM, further data preprocessing is not needed as the model itself can eliminate unrelated textual data. In addition, retaining emojis and other non-textual elements can also be retained and might even help to reveal the sentiment of the writer.

B. MODEL SELECTION

We selected a range of LLMs and more traditional NLP methods to evaluate the usefulness of LLMs for SQ assessment. For LLMs, we opted to use *ChatGPT 3.5* by OpenAI [53] as it is the most recent freely available version of this tool, which has previously been shown to have a beneficial impact on organizational performance [5]. We also included a current competitor to compare our results and explore the potential of LLMs beyond ChatGPT.

We selected *Claude 3* by AI Unicron Anthropic, whose developers claim it can outperform OpenAI's LLM [54]. The Claude series is a group of LLMs using generative pre-trained transformer (GPT) architectures. Initiated in March 2023 and further expanded with Claude 3 in March 2024, this series

[2]https://www.kaggle.com/datasets/prakharrathi25/google-play-store-reviews

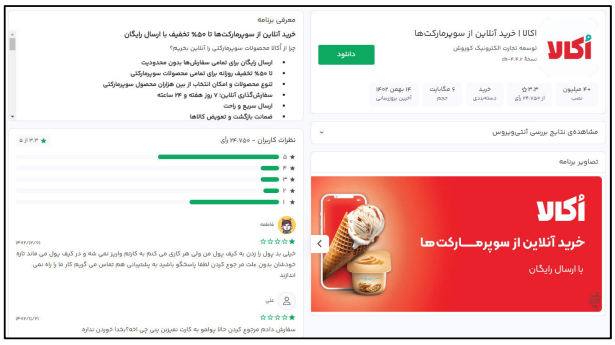

FIGURE 2. Cafebazaar website.

introduces capabilities for analyzing text and visual data [55]. We chose to use these already trained models (over models specifically trained for a determined purpose such as sentiment analysis in our case), to assess further how LLMs can be used in a practical scenario (e.g., by domain experts in organizations who want to augment their work using ChatGPT's capabilities). In addition, the APIs for these LLMs are compatible with widely used programming languages like Python, further simplifying their accessibility.

To have a further comparison of the results that can be created using the two selected LLMs, we also applied some more traditional text processing models. For this purpose, we applied three additional models to the second step in our methodology (Sentiment Analysis). We applied *TextBlob*, a Python library for natural language processing tasks [56], which uses the Naive Bayes classifier and probabilistic models [57]. *VADER*(Valence Aware Dictionary and sEntiment Reasoner), a rule-based sentiment analysis tool designed specifically for analyzing sentiment in text using the Lexicon approach [58], and *Transformers*, a type of deep learning model processing, which utilizes attention mechanisms to capture relationships between words in a sequence [59].

C. PROMPT DEFINITION

Choosing the right prompt for an LLM is critical, as it directly impacts the quality of the model's responses. To ensure the efficacy of the prompt, we conducted tests with various prompts tailored to each specific use case (i.e., identification of SQ-related content and sentiment of the review). Our approach involved drawing examples from Zhang et al. [60], adhering to the OpenAI API [61] guideline, and emphasizing the formulation of clear and straightforward prompts.

Our prompts consist of two distinct components. The first part is the command or query used to interact with the model and elicit a response. In specific use cases, such as categorizing comments and performing sentiment analysis, the second part involves the data the model needs to process to generate the desired outcome. This data is transformed into a JSON string. We provide an example of such an interaction below:

Here are customer reviews for an Iranian Omnichannel retailer in the given json string. Based on the review answer following questions in English.

In one word what type of service quality dimension has been mentioned in the text?

What is the sentiment of the text based on these three classes: positive, neutral, or negative?

Print your result in a json string including 'ReviewId', 'Channel', 'Service_Quality_Dimension', 'Main_Concern', 'Sentiment', 'Satisfaction_Source'. Just print a json string nothing else.

[{"Id":75552883,"Review":"فوق العاده ست واقعا دارن زحمت
به کوری چشششششششم "Id":76792157,"Review":"}میکشند "},
ایراد "Id":71481956,"Review":"}بعضیا م برای ی "},
فروشگاهیه که توتموم گرونی "Id":74217171,"Review":"}داااااره"},
کار نمی "Id":71271524,"Review":"}پناه مردم هست"},
{"Id":70530031,"Review":"}ایرادات دارد باز"},
"Id":68768897,"Review":"}متاسفانه نرمافزار مناسب ی نیست"},
واقعاً بی کاربرده"},{"Id":63169771,"Review":"
{"Id":60818080,"Review":"هیچ کارایی ندارد"}]

Our process starts by reading reviews from an *Excel* file into a *Pandas* dataframe. We then use a crafted prompt to instruct Claude and ChatGPT to categorize the feedback. Through Python, we orchestrate batch requests to the Anthropic and OpenAI APIs, sending groups of 10 reviews at a time and receiving insights in JSON format. This systematic approach enables efficient processing and aggregation of data, highlighting the integration of cutting-edge AI with practical data analysis techniques to extract actionable insights from textual feedback.

The sentiment analysis process with the three selected traditional models starts with importing the necessary Python libraries and modules. In our case, we have utilized the Natural Language Toolkit (NLTK), specifically the VADER sentiment analysis tool, TextBlob, and Pandas for handling data operations. For VADER, the 'vader_lexicon' has been downloaded to access sentiment based on lexical features. The CSV file has been imported into Python; this file contains the text data (reviews or comments) to be analyzed. Initializing the Sentiment Analyzer involves creating an instance of SentimentIntensityAnalyzer. This tool computes sentiment scores for the text data based on the intensity of both positive and negative words. The compound score is then used to classify the overall sentiment as 'positive,' 'negative,' or 'neutral.' The sentiment classification and the compound score are stored in new columns within the DataFrame by splitting the returned tuple from the function into two separate columns. Sentiment analysis with Transformers differs slightly as it involves utilizing the 'pipeline' and 'DistilBertTokenizer' from the Transformers library to handle sentiment analysis and tokenization. The process begins by setting up the sentiment analysis pipeline using the 'pipeline' function from the Transformers library, which is specifically designed for sentiment analysis tasks. Then, the 'DistilBertTokenizer' is loaded to handle text preprocessing. This involves a function called 'analyze_sentiment_transformers' that takes a string of text as input. Within this function, the tokenizer is used to prepare the text for the model by truncating it to a maximum length (512 tokens in this case) and converting it to the format required by the model. After preprocessing, the text is passed to the sentiment analysis pipeline to predict the sentiment and corresponding confidence score.

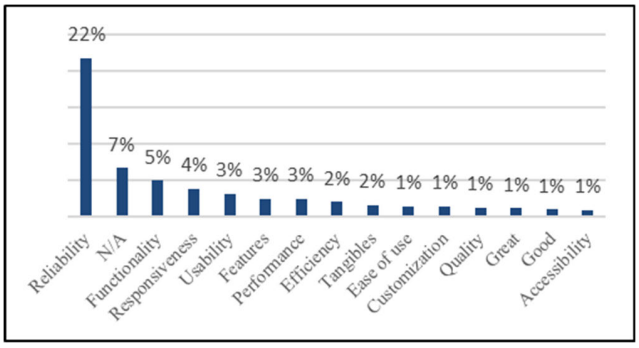

FIGURE 3. The most common SQ dimension extracted by ChatGPT (dataset 1).

D. EVALUATION

We finally put in place routines to assess the quality of the output that is created by the employed LLMs and the more traditional text processing models to ensure the reliability of the created results, before their actual content is interpreted and compared in more detail.

For the first step (*Extraction of SQ-related content*), we assessed the type of SQ dimension extracted by the employed methods. This was first done on the complete set of samples for both datasets to see what the result would look like if the LLMs were used without further guidance. Then, a subsample of reviews was randomly drawn from both datasets (the Persian reviews drawn were translated to English) and assessed by human reviewers based on a given set of dimensions (detailed in the RESULTS Section). Two researchers independently assessed the reviews, and to avoid personal bias, only instances where both reviewers agreed on the identified SQ dimension have been taken into consideration. This was done until a sample of 450 reviews was reached, of which 87 were subsequently excluded because both reviewers agreed that they were not related to SQ, which resulted in a final sample of 363 reviews. These reviews were then used to assess how reliably both LLMs could assess whether a review was related to SQ and which dimension of SQ it was mainly related to.

For the second step (*Sentiment analysis*), the same sub-sample of reviews related to SQ ($n = 363$) was used and its sentiment (negative, neutral, positive) was assessed by the two same researchers. The resulting groundtruth sample (based on human judgment) was then used to assess the

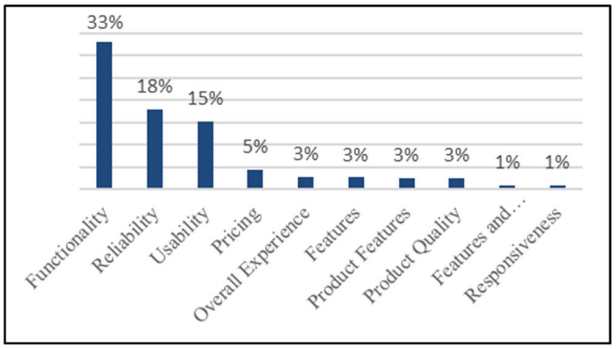

FIGURE 4. The most common SQ dimension extracted by Claude 3 (dataset 1).

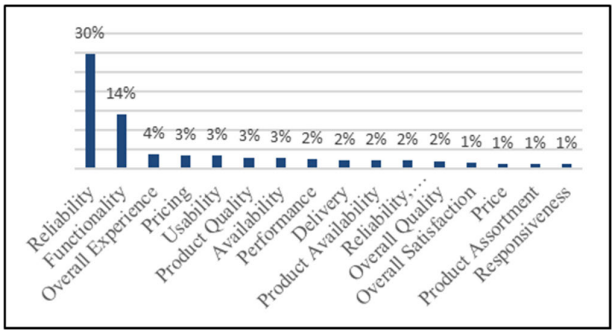

FIGURE 6. The most common SQ dimension extracted by Claude 3 (dataset 2).

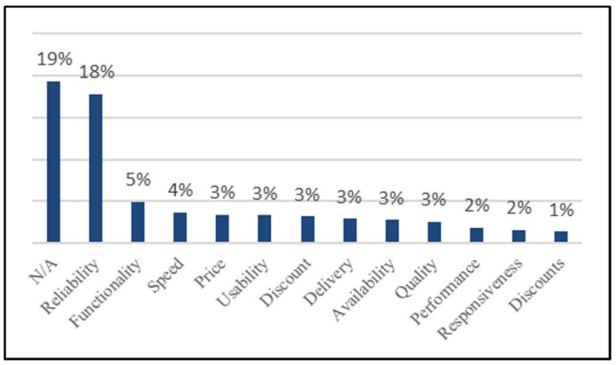

FIGURE 5. The most common SQ dimension extracted by ChatGPT (dataset 2).

interrater agreement of each text mining method with human raters based on Cohen's Kappa [62], [63].

IV. RESULTS

A. EXTRACTION OF SQ-RELATED CONTENT

For dataset 1, running the prompt "In one word, what type of service quality dimension has been mentioned in the text?" resulted in 171 different outcomes from ChatGPT and 148 different outcomes from Claude 3. In the case of ChatGPT, the 15 most common outcomes (i.e., applied to at least 1% of the reviews each) covered roughly 60% of the cases (see Figure 3), whereas for Claude 3, the ten most common outcomes (i.e., applied to at least 1% of the reviews each) covered more than 80% of the cases (see Figure 4). For this dataset, only around 7% of cases were not classified by ChatGPT (i.e., resulting in "N/A").

For dataset 2, running the prompt "In one word, what type of service quality dimension has been mentioned in the text?" resulted in 580 different outcomes from ChatGPT and 338 different outcomes from Claude 3. In the case of ChatGPT, the 13 most common outcomes (i.e., applied to at least 1% of the reviews each) covered more than 60% of the cases (see Figure 5), whereas for Claude 3, the 16 most common outcomes (i.e., applied to at least 1% of the reviews each) covered more than 70% of the cases (see Figure 6). The results differ substantially in some regards, including a high level of reviews that could not be classified into SQ dimensions by ChatGPT (i.e., "N/A" the most common outcome with 19.28% of all cases) compared to just under 1% of such cases for Claude 3.

Further observations can be made regarding the content and phrasing of the dimensions provided by both LLMs. In contrast, many comments have been correctly classified into well-known service quality dimensions, and numerous others have been categorized into AI-generated groups without any contextual background (e.g., "decoration" rather than "physical aspects" or "speed" instead of "reliability"). Similarly, although we have asked the AI to provide a summary of the service quality in one word, in some mixed cases, Claude 3 extracted a combination of dimensions, for instance, "*You made your goods more expensive. Also, the cost of its transportation. The quality of the meat is also good. He did not bring me a kilo of beans, you did not follow up at all*" Claude 3 extracted [Product Quality, Pricing, Customer Service]. The comparison table is provided in APPENDIX B.

Aside from these more qualitative observations that can give us an initial overview of the types of outputs the LLMs can provide, we also tried to compare their results more systematically. For this purpose, we inspected a sub-sample of n = 363 reviews (as mentioned in EVALUATION section) and the dimensions that were assigned to them by both LLMs. First, though, we also looked at the 87 reviews that had to be removed initially as they were unrelated to SQ. The goal was to identify which of the different strategies by both LLMs regarding handling non-SQ-related content was more accurate. ChatGPT assigned a "N/A" label to a larger portion of reviews (e.g., 22% in dataset 1). At the same time, Claude barely uses this approach and instead assigns a label to almost every review. For instance, for those that we could not allocate any SQ, the outcome of LLMs is shown in TABLE 1 For the Persian Dataset.

We then investigated whether the resulting dimensions show a fit with prominent theoretical models of SQ and can, therefore, be compared to the outputs generated by more environment standardized methods such as questionnaires. To find appropriate theoretical SQ dimensions, we referred to the sources discussed in the SERVICE QUALITY (SQ)

TABLE 1. Undefinable SQ sample.

SQ Dimension	ChatGPT	Claude dim
#N/A	59	5
Policies	12	23
Online Service	3	23
Reliability	10	7
Personal Interaction	2	
overall service		33

ASSESSMENT section and found that the CALSUPER model [19] fits closest to the more specific case of supermarket retail that is represented by our second dataset. The dimensions included in this model are *physical aspects*, which refer to the physical aspects of the service environment, facilities, and equipment; *reliability*, which refers to the ability to deliver promised services dependably, accurately, and consistently; *personal interaction*, which refers to the direct communication and engagement between service providers and customers throughout the service delivery; and *policies*, which refers to the consistency and fairness of a service provider's rules, regulations, and procedures. As the supermarket represented in dataset 2 offers a mobile app and mobile loyalty app, some application-related comments could not be allocated to any other dimension. In response, we have added a new dimension called *online service*. We also created an 'overall service' dimension to capture comments where the customer mentioned the service in general terms without specifying any particular aspects of it and simply expressed a positive or negative opinion.

To assess the similarity between the outcomes generated by the LLMs and those found in the literature, we utilized the Jaccard Similarity test [64], represented by formula 1. Here, the greater extent of dimensions generated based on dataset 2 (580 for ChatGPT and 338 for Claude 3 – see also the Supplementary Material for a full list of all dimensions created by both LLMs for both datasets) was used as the input and compared to the dimensions of the chosen theoretical model. The similarity result is less than two percent for both methods, indicating a significant difference between the initial outcomes of ChatGPT and the literature.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

In the next step, we abstracted the dimensions that can be mapped to more general theoretical models of SQ. For this purpose, we had to allocate the outcomes of the LLMs to the SQ, as mentioned above dimensions. Two independent researchers did This allocation manually and assigned the outputs for dataset 2 generated by both LLMs to the chosen theoretical SQ dimensions. In different evaluations, the results were discussed until a consensus was reached. Exemplary phrases used by the two LLMs for each SQ dimension are presented in TABLE 2.

TABLE 2. Common phrases used by LLMS for each SQ dimension.

Dimension	Common phrases for ChatGPT	Common phrases for Claude 3
Physical aspects	Tangibles, location, Decoration, design, store environment, Appearance, Aesthetics, Environmental Sustainability	Convenience for people, Ease of access, coverage, Appearance
Reliability	Reliability, Speed, Usability, Performance, Accuracy, Product Availability, Efficiency, Service Delivery, Assurance, Security, Stock availability, Serviceability,	Slow Speed, Slow performance, Product availability, Delayed Delivery , Cost-effectiveness, available products Timeliness of delivery, out of stock Difficulty in purchase, Ease of use Delay in-order delivery, Speed, Accuracy Unavailability, Technical Issues, Unclear, confusing
Personal interaction	Communication, Empathy, Staff, Relationship, communication with clients, Behavior, Store support, Patience, Emotional experience, Culture,	polite and respectful service, team's efforts employees, hardworking staff, young staff, Appreciation for non-physical services Professionalism. Helpful staff
Policies	Price, Discount, Quality, Support, Variety, Promotion, Cost, the Payment, Product Range, Discounts and promotions, Coupon	Discount, product variety, product selection, delivery fee, delivery charges, promotions, Affordability
Online service	Timeliness, Website, App, software, Installation, technical quality, Application function, Authentication, User experience, Technology, Verification, Website/App, Update, Application Performance	app working, Application Performance, Application Quality, app practical, app loading, rooted devices, version of the app, Issue applying discount code, verification codes, Location verification, Location service, Internet Connection,
Overall Service	Overall experience, Overall satisfaction, Excellence, General, Overall quality, Overall service quality, overall_poor_service,	Dissatisfaction with service, General Feedback, Poor quality or performance Negative Experience, General Impression , Positive Experience, General Satisfaction Poor Quality, Recommendation

This mapping was then used to analyze the SQ dimensions that were assigned to each of the previously labeled 363 reviews. The resulting distribution of the mentioned SQ dimensions is shown in Figure 7. Based on this mapping and the comparison with the labels of human raters, we find that 76% of reviews have dimensions classified correctly by ChatGPT. In comparison, Claude 3 could identify dimensions with 68% accuracy. In this context, we observed some differences that are related to short comments where Claude 3 assigns the *Overall Service* dimension, which is not absolutely wrong

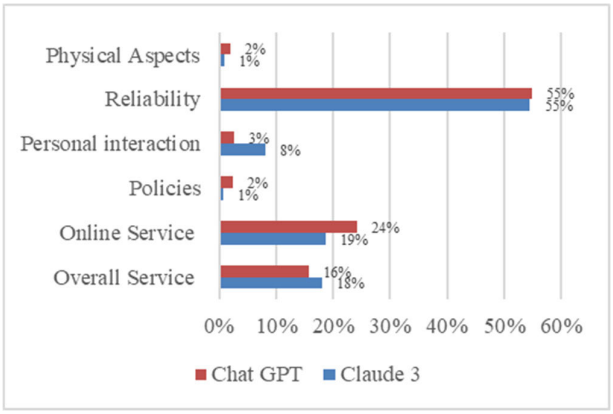

FIGURE 7. Service quality dimension allocation.

TABLE 3. Interrater agreement for human raters and text mining methods.

	Human raters	Chat GPT	Clau de 3	Text Blob	Transf ormer	Va der
Negative	209	236	251	88	249	106
Neutral	77	53	46	137	-	137
Positive	77	74	66	138	114	120
Kappa index – dataset 2	-	.64	.57	.28	.51	.28
Negative	213	233	256	95	205	169
Neutral	76	55	46	140	-	101
Positive	74	75	61	128	158	104
Kappa index – dataset 1	-	.67	.59	.27	.42	.30

but may not add significant value. A further comparison of the distribution of the assigned dimensions based on the results depicted in Figure 7 Using the Kolmogorov-Smirnov (KS) does not show any significant difference ($P = 0.93$), which does not indicate any further more obvious systematic differences in the results provided by both LLMs at this point.

B. SENTIMENT ANALYSIS

For the second step in our analysis, we again used the previously mentioned sub-sample of 363 reviews that deal with SQ, which have also been rated for their sentiment (negative, neutral, positive) by two independent researchers. In case of differing views, the results were again discussed until a consensus was reached. In addition to the two LLMs, we included the results of three sentiment analysis methods: TextBlob [56], VADER [58], and a neural network-based approach using transformers provided via the Hugging Face libraries [59]. The first drawback of these libraries was the lack of support for the Persian language. Consequently, we had to translate the Persian language comments into English.[3] After translating all the comments, we fed the Python script with the translated outcomes.

In the subsequent part of the prompt, the LLMs were tasked to assess the sentiment (negative, neutral, positive) in customer reviews for a retail store. The output of the more traditional NLP methods and the LLMs were then compared to the human judgments. For this, we used Cohen's Kappa statistic. The Cohen's Kappa index metric shows agreement levels between the models' outcomes and expert reviews [62], [63]. In TABLE 3, the overall classifications (i.e., for all 363 reviews, independent of which dataset they belong to; examples for reviews and their evaluation by each method are provided in Appendix A), and the Kappa scores for each data of each method and dataset are shown. We can see that whereas all methods achieve a performance greater than chance alone, the LLMs outperform the more traditional methods.

[3]For this purpose, we used www.onlinedoctranslator.com/en/.

Only the employed *Transformer* came close to the performance of the LLMs on the second dataset. Based on the categorization of the strength of agreement by Landis and Koch [62], we can add that ChatGPT achieves substantial agreement on both datasets, and Claude 3 achieves moderate agreement on both datasets. The remaining methods achieve fair to moderate agreements. A further interesting observation that can be made is that the performance on both datasets is roughly equivalent. Though also the Persian reviews (dataset 2) had to be translated to English for traditional models, further adaptations were not made, and therefore, additional differences that could stem from different cultural customs, manners of expression, or other contextual differences did not substantially impact the performance of employed methods.

To finally demonstrate the type of output that managers could expect in practice from applying an LLM for the assessment of SQ, we then also applied the previously generated mapping of LLM output dimensions to theoretical SQ dimensions to all reviews in dataset 2. We chose dataset 2 due to the greater extent of dimensions created by both LLMs for this dataset. We focused solely on the classifications by ChatGPT as it was shown to perform better both in the extraction of SQ dimensions (76% of correct labels compared to 68% by Claude) and the analysis of sentiment (Kappa index of.64 on dataset 2 compared to.57 by Claude 3). To analyze the overall sentiment for each SQ dimension, we created a sentiment score for each dimension, assigning a score of 1 to positive comments, −1 to negative comments, and 0 to neutral comments. The result is displayed in TABLE 4.

V. DISCUSSION

This study dealt with the question of how current LLMs can contribute to handling large amounts of user-generated content in the context of SQ assessment. Whereas we find that the final result of applying such models can look comparable to what we might expect from the application of more standardized methods like questionnaires (see TABLE 3), from a practical point of view, we have to consider what effort was put into this process and what utility it provides to potential users.

TABLE 4. Sentiment analysis for SQ dimensions based on ChatGPT and dataset 2.

SQ dimension	Negative	Neutral	Positive	Sentiment Score
Physical aspects	119	6	16	-0.73
Reliability	1206	96	192	-0.68
Personal interaction	71	9	33	-0.34
Policies	2777	216	377	-0.71
Online Service	889	37	48	-0.86
Overall service	96	11	56	-0.25
#N/A	759	375	444	-0.20

In general, we tried to apply a text mining approach that is as generic as possible to see how the employed LLMs perform without further specialized prompting or feedback loops to perhaps explain and improve their results. Still, for the first step in the SQ assessment process (i.e., dimension extraction), arguably the more complex step, we find that the LLMs can deal with the largely unstructured UGC that was acceptably provided. Here, we must consider, though, that the generated SQ dimensions may not entirely fit the expectations of human users. Therefore, in our case, an intermediate step was needed to map the created dimensions to SQ dimensions established in extensive research. Yet, even the worst case that we had to deal with (i.e., 580 outcome dimensions by ChatGPT on dataset 2) is still substantially smaller in magnitude if compared to the size of the original dataset (i.e., more than 9,000 reviews). Hence, a reduction in effort can be achieved. In addition, this mapping can provide us with further information on the important topics within each dimension (see, for example). TABLE 2), which generates further insights into the issues that may drive SQ-related reviews and goes above and beyond what could be an expected output from a standardized questionnaire.

A. EXTRACTION OF SQ-RELATED DIMENSIONS

For SQ dimension extraction specifically, we have to be conscious of potential problems that may impact the accuracy of the employed LLMs. As our evaluation on a sub-sample of reviews showed, both ChatGPT, with 76%, and Claude 3, with 68%, achieve commendable accuracy in labeling the reviews even without providing them with a classification system, which is in line with results in previous research [4]. Although we did not find any systematic differences between both LLMs in terms of their overall results, there are nonetheless differences that will be important to practitioners. First, it will be important for practitioners to be able to initially assert whether reviews used for SQ assessment reference any topics related to SQ. Therefore, the chosen method should be able to indicate whether a review is not relevant to the task. In this regard, the two chosen LLMs show different approaches, with ChatGPT not classifying reviews if they seemingly cannot be labeled with an SQ dimension, whereas Claude 3 still gave reviews a label. Through analysis of a sub-sample evaluated by human raters, we found that ChatGPT performs better in identifying non-relevant reviews, with 68% of the comments that cannot be allocated from the expert view being identified as #N/A by it in contrast less than around 5% of them have been allocated correctly by Claude 3 and near 50% of them have been assigned to Overall service correctly.

Second, though we tried to somewhat force the assignment of a specific label by purposefully prompting the LLMs to assign only one SQ dimension to each review, which ChatGPT adhered to, consumer reviews can simultaneously address more than one SQ-related issue. Accordingly, rather than using the closest fitting SQ dimension as a label, which was the strategy chosen by ChatGPT, Claude 3 resorted to creating labels that combined multiple dimensions, reflecting the ambiguity of the multitude of topics potentially contained in one review. Here, it depends on the use case and further processing of the results, which approach practitioners should prefer. Whereas the approach chosen by ChatGPT allows for faster data processing, as one label can be assumed for each review, the approach by Claude 3 may closer reflect the actual content of a review if multiple topics are addressed frequently within consumer reviews. In their current state, though, ChatGPT still performed slightly better in identifying the best fitting SQ dimension (76% compared to 68% by Claude 3) if that is the approach to be chosen by practitioners.

B. SENTIMENT ANALYSIS

When identifying the sentiment expressed in a review, both LLMs substantially outperformed all the employed more traditional NLP methods, both on a dataset initially in English and one originally in Persian. In particular, ChatGPT showed substantial agreement with two human raters who had evaluated a subset of reviews to create a ground truth against which the performance of the models could be compared. This finding aligns with other recent studies, such as Fatouros et al. [47] or Leippold [40], which also argued that LLMs show reliable performance on standardized text analysis tasks such as sentiment analysis. An additional observation that can be made if the results of this task are compared to the results of the previous task is that when labels are already provided. Hence, the task is further limited in its scope (i.e., in the first task, the LLMs not only had to assign a label for an SQ dimension but also come up with the dimensions on their own), the expected performance can already be satisfactory without the need for any intermediate steps (e.g., the mapping we implemented to fit the generated SQ dimensions into a theoretical model).

Overall, the results are intriguing and have potential utility in the business domain. However, out of the box, that is, without further specific training and adaptations of the model to a given task, there is still a need for human supervision. For example, consider that both LLMs labeled a review that contained the phrase ''Dan Birthday Oct 28'' with the SQ dimension *Reliability.* As it is not apparent what led to this label, further inquiries of the LLM are needed to extract

the reasoning behind it. In turn, potential adaptations of the prompting scheme might be needed to enhance the accuracy of the SQ assessment. Hence, whereas we can conclude that LLMs can undoubtedly help to streamline the process of SQ assessment, in particular, if we want to make use of the enormous extent of UGC that is now available, further tuning is needed to turn a publicly available LLM into a reliable means to automate a greater extent of this process.

C. LIMITATIONS AND FURTHER RESEARCH

As we tried to apply an approach to SQ assessment that is not specific to a given dataset or a theoretical model, we exhibited limitations in the resulting accuracies that could potentially be overcome with further adaptations of the methodology. For example, mentioned by Alexander et al. [44], incorporating domain knowledge into text mining approaches is an important step to ensure their quality. In the case of the specific tasks covered as part of this study, domain knowledge could, for example, be used to construct the set of SQ dimensions that the LLM subsequently uses to label reviews or, if practitioners want not to restrict themselves to just a set of dimensions, domain knowledge could play a role when constructing a mapping that is then provided to the LLM again to reanalyze the data. In addition, further prompting could be used to explore the reasoning behind the assignment of a label (e.g., which word combinations most strongly influenced the assignment), which could also provide the domain experts with further insights into topics that might also be important for a specific SQ dimension outside of what they might have considered. As the study reported here was supposed to be an initial test of the utility of LLMs for SQ assessment with potentially as few adaptations made by practitioners as possible, as they might not be familiar with the mechanisms behind an LLM, such further adaptations of the prompting approach, in particular, are subject for future research (e.g., refer to Han et al. [65] for an approach to sentiment analysis with ChatGPT that provides even better accuracy than the simple approach reported here).

Aside from the methodology employed by the user of an LLM, it should also be noted that we employed widely available, multi-purpose LLMs that may not be ideally suited for each task in this study. For example, Rostami et al. [66] report on the development of the model for information extraction specifically suited for Persian and English datasets, which might exhibit better accuracies in the specific use cases that were the basis for this study or consider Wei et al. [67] who created a system based on ChatGPT that could also provide better accuracies on information extraction tasks than a basic version of ChatGPT.

VI. CONCLUSION

The widespread introduction of LLMs has marked a fundamental paradigm shift in data preprocessing and coding procedures. Now, with prompt management, individuals can efficiently execute models quickly. Embracing this change can positively impact management's overall quality of data

TABLE 5. Example of review sentiment.

Content	ChatGPT	CLAUDI3	TEXTBLOB	TRANSFORMERS	VADER
I cannot open the app anymore.	Neg.	Neg.	Neu.	Neg.	Neu.
I have been begging for a refund from this app for over a month and nobody is replying me.	Neg.	Neg.	Neu.	Neg.	Neu.
Very costly for the premium version (approx Indian Rupees 910 per year). Better to download the premium version of this app from apkmos website and use it. Microsoft to do list app is far more better.	Neg.	Neg.	Pos.	Neg.	Pos.
Used to keep me organized, but all the 2020 UPDATES have made a mess of things !!! Y cudn't u leave well enuf alone ??? Guess ur techies feel the need to keep making changes to justify continuing to collect their salary !!! ðŸ¤¬ðŸ¤¬ðŸ¤¬	Neg.	Neg.	Neg.	Neg.	Neg.
It has changed how I viewed my different lists. Now they are all jumbled together and I can't find what I need.	Neg.	Neg.	Neu.	Neg.	Neu.
I'm only looking for a grocery list app but every time I tap away from the app I have to tap again (after opening it again) to see the list. I can't find a way to keep a certain list showing when I open or reopen the app. eta: in response to the reply, it doesn't work like that on my phone. Even if the grocery list is showing, when I open another app and then go back to the any.do app, the list of lists is showing and I have to retap to get the grocery list to show again.	Neg.	Neg.	Pos.	Neg.	Neu.
Reset my free trial, new phone I'd like to see if it's better.	Neu.	Neu.	Pos.	Neg.	Pos.
How do to stop monthly payment because i don't use this app anymore	Neg.	Neu.	Neu.	Neg.	Neg.

TABLE 6. Dimension comparison.

Dimension	ChatGPT Dataset1	Claud 3 Dataset1	ChatGPT Dataset2	Claude Dataset2
Reliability	0.22	0.18	0.18	0.30
Functionality	0.05	0.33	0.05	0.14
N/A	0.07		0.19	
Usability	0.03	0.15	0.03	0.03
Pricing		0.05		0.03
Responsiveness	0.04	0.01	0.02	0.01
Performance	0.03		0.02	0.02
Overall Experience		0.03		0.04
Features	0.03	0.03		
Availability			0.03	0.03
Product Quality		0.03		0.03
Delivery			0.03	0.02
Price			0.03	0.01
Speed			0.04	
Quality	0.01		0.03	
Discount			0.03	
Product Features		0.03		
Product Availability				0.02
Reliability, Responsiveness				0.02
Efficiency	0.02			
Tangibles	0.02			
Overall Quality				0.02
Discounts			0.01	
Overall Satisfaction				0.01
Product Assortment				0.01
Ease of Use	0.01			
Customization	0.01			
Great	0.01			
Good	0.01			
Accessibility	0.01			
Features and Functionality		0.01		

utilization. We have investigated the capability of Large Language Models for a real-world business application, specifically focusing on tasks like topic modeling and sentiment analysis in the context of SQ assessment. Our choice of LLMs was ChatGPT and Claude 3, a cutting-edge model built on the transformer architecture. Upon evaluating the LLMs' performance, our findings indicate they can attain commendable accuracy and quality, showcasing their practicality and adaptability for real-world data applications. Additionally, our study illustrates that ChatGPT exhibits proficiency in managing diverse languages and data formats, including Persian.

APPENDIX A
SAMPLE OF SENTIMENT ANALYSIS RESULTS

See Table 5.

APPENDIX B
MOST COMMON DIMENSION COMPARISON

See Table 6.

REFERENCES

[1] L. Lu, P. Xu, Y.-Y. Wang, and Y. Wang, "Measuring service quality with text analytics: Considering both importance and performance of consumer opinions on social and non-social online platforms," *J. Bus. Res.*, vol. 169, Dec. 2023, Art. no. 114298, doi: 10.1016/j.jbusres.2023.114298.

[2] A. Ūsas, E. Jasinskas, and D. Štreimikienė, "The influence of websites quality on users e-loyalty in the online store," *Polish J. Manage. Stud.*, vol. 28, no. 1, pp. 344–359, 2023, doi: 10.17512/pjms.2023.28.1.20. [Online]. Available: https://pjms.zim.pcz.pl/resources/html/article/details?id=617132&language=en

[3] M. Holmlund, Y. Van Vaerenbergh, R. Ciuchita, A. Ravald, P. Sarantopoulos, F. V. Ordenes, and M. Zaki, "Customer experience management in the age of big data analytics: A strategic framework," *J. Bus. Res.*, vol. 116, pp. 356–365, Aug. 2020, doi: 10.1016/j.jbusres.2020.01.022.

[4] T. Brown et al., "Language Models Are Few-Shot Learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[5] M.-N. Chu, "Assessing the benefits of ChatGPT for business: An empirical study on organizational performance," *IEEE Access*, vol. 11, pp. 76427–76436, 2023, doi: 10.1109/ACCESS.2023.3297447.

[6] W. Duan, Q. Cao, Y. Yu, and S. Levy, "Mining online user-generated content: Using sentiment analysis technique to study hotel service quality," in *Proc. 46th Hawaii Int. Conf. Syst. Sci.*, Mar. 2013, pp. 3119–3128.

[7] J. Mejia, S. Mankad, and A. Gopal, "Service quality using text mining: Measurement and consequences," *Manuf. Service Operations Manage.*, vol. 23, no. 6, pp. 1354–1372, Nov. 2021, doi: 10.1287/msom.2020.0883.

[8] M. U. Hadi, Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili. *A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage*. Accessed: Jan. 25, 2024. [Online]. Available: https://www.techrxiv.org

[9] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit. Health*, vol. 2, no. 2, Feb. 2023, Art. no. e0000198, doi: 10.1371/journal.pdig.0000198.

[10] M. Kosinski, "Theory of mind may have spontaneously emerged in large language models," 2023, *arXiv:2302.02083*.

[11] F. Xing, "Designing heterogeneous LLM agents for financial sentiment analysis," 2024, *arXiv:2401.05799*.

[12] *Open AI & GPT News, 92% of Fortune 500, 100M Using ChatGPT Weekly*. Accessed: Jan. 25, 2024. [Online]. Available: https://www.linkedin.com/pulse/92-fortune-500-100m-using-chatgpt-weekly-open-ai-gpt-news-pietc/

[13] R. W. Puyt and D. Ø. Madsen, "Evaluating ChatGPT-4's historical accuracy: A case study on the origins of SWOT analysis," *Frontiers Artif. Intell.*, vol. 7, May 2024, Art. no. 1402047, doi: 10.3389/frai.2024.1402047.

[14] N. Donthu, D. D. Gremler, S. Kumar, and D. Pattnaik, "Mapping of journal of service research themes: A 22-year review," *J. Service Res.*, vol. 25, no. 2, pp. 187–193, May 2022, doi: 10.1177/1094670520977672.

[15] D.-G. Oh and B. Elango, "Quantitative analysis of the publication trends in four decades of SERVQUAL research: A bibliometric approach," *Total Quality Manage. Bus. Excellence*, vol. 35, nos. 1–2, pp. 297–319, Jan. 2024, doi: 10.1080/14783363.2023.2293877.

[16] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "A conceptual model of service quality and its implications for future research," *J. Marketing*, vol. 49, no. 4, pp. 41–50, Sep. 1985, doi: 10.1177/002224298504900403.

[17] J. J. Cronin and S. A. Taylor, "Measuring service quality: A reexamination and extension," *J. Marketing*, vol. 56, no. 3, p. 55, Jul. 1992.

[18] P. A. Dabholkar, D. I. Thorpe, and J. O. Rentz, "A measure of service quality for retail stores: Scale development and validation," *J. Acad. Marketing Sci.*, vol. 24, no. 1, pp. 3–16, Dec. 1996, doi: 10.1007/bf02893933.

[19] R. Vázquez, I. A. R.-D. Bosque, A. Ma Díaz, and A. V. Ruiz, "Service quality in supermarket retailing: Identifying critical service experiences," *J. Retailing Consum. Services*, vol. 8, no. 1, pp. 1–14, Jan. 2001, doi: 10.1016/s0969-6989(99)00018-1.

[20] S. Akinci, E. Atilgan-Inan, and S. Aksoy, "Re-assessment of E-S-qual and E-RecS-qual in a pure service setting," *J. Bus. Res.*, vol. 63, no. 3, pp. 232–240, Mar. 2010, doi: 10.1016/j.jbusres.2009.02.018.

[21] A. Parasuraman, V. A. Zeithaml, and A. Malhotra, "E-S-QUAL: A multiple-item scale for assessing electronic service quality," *J. Service Res.*, vol. 7, no. 3, pp. 213–233, Feb. 2005, doi: 10.1177/1094670504271156.

[22] X. J. Mamakou, P. Zaharias, and M. Milesi, "Measuring customer satisfaction in electronic commerce: The impact of e-service quality and user experience," *Int. J. Quality Rel. Manage.*, vol. 41, no. 3, pp. 915–943, Feb. 2024, doi: 10.1108/ijqrm-07-2021-0215.

[23] R. Zhang, M. Jun, and S. Palacios, "M-shopping service quality dimensions and their effects on customer trust and loyalty: An empirical study," *Int. J. Quality Rel. Manage.*, vol. 40, no. 1, pp. 169–191, Jan. 2023, doi: 10.1108/ijqrm-11-2020-0374.

[24] D. A. Rashid and D. R. Rasheed, "Logistics service quality and product satisfaction in e-commerce," *SAGE Open*, vol. 14, no. 1, 2024, doi: 10.1177/21582440231224250.

[25] J. Krumm, N. Davies, and C. Narayanaswami, "User-generated content," *IEEE Pervasive Comput.*, vol. 7, no. 4, pp. 10–11, Oct. 2008, doi: 10.1109/MPRV.2008.85.

[26] L. Hagen, "Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models?" *Inf. Process. Manage.*, vol. 54, no. 6, pp. 1292–1307, Nov. 2018, doi: 10.1016/j.ipm.2018.05.006.

[27] Z. Yang and X. Fang, "Online service quality dimensions and their relationships with satisfaction," *Int. J. Service Ind. Manage.*, vol. 15, no. 3, pp. 302–326, Jul. 2004, doi: 10.1108/09564230410540953.

[28] N. Korfiatis, P. Stamolampros, P. Kourouthanassis, and V. Sagiadinos, "Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews," *Expert Syst. Appl.*, vol. 116, pp. 472–486, Feb. 2019, doi: 10.1016/j.eswa.2018.09.037.

[29] I. Sutherland and K. Kiatkawsin, "Determinants of guest experience in airbnb: A topic modeling approach using LDA," *Sustainability*, vol. 12, no. 8, p. 3402, Apr. 2020, doi: 10.3390/su12083402.

[30] M. Alzate, M. Arce-Urriza, and J. Cebollada, "Mining the text of online consumer reviews to analyze brand image and brand positioning," *J. Retailing Consum. Services*, vol. 67, Jul. 2022, Art. no. 102989, doi: 10.1016/j.jretconser.2022.102989.

[31] L. Yao, C. Mao, and Y. Luo, "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks," *BMC Med. Informat. Decis. Making*, vol. 19, no. S3, p. 71, Apr. 2019, doi: 10.1186/s12911-019-0781-4.

[32] M. Garofalakis, R. Rastogi, and K. Shim, "Mining sequential patterns with regular expression constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 3, pp. 530–552, May 2002, doi: 10.1109/TKDE.2002.1000341.

[33] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Inf. Process. Manage.*, vol. 39, no. 1, pp. 45–65, Jan. 2003, doi: 10.1016/s0306-4573(02)00021-3.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, vol. 1, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[35] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[37] R. Alec, W. Jeffrey, C. Rewon, L. David, A. Dario, and S. Ilya, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, Feb. 2019. Accessed: Jan. 26, 2024. [Online]. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[38] J. Pedro, I. Brown, and M. Hart, "Capabilities and readiness for big data analytics," *Proc. Comput. Sci.*, vol. 164, pp. 3–10, Jan. 2019, doi: 10.1016/j.procs.2019.12.147.

[39] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Dec. 2022, pp. 22199–22213.

[40] M. Leippold, "Thus spoke GPT-3: Interviewing a large-language model on climate finance," *Finance Res. Lett.*, vol. 53, May 2023, Art. no. 103617, doi: 10.1016/j.frl.2022.103617.

[41] V. Browning, K. K. F. So, and B. Sparks, "The influence of online reviews on consumers' attributions of service quality and control for service standards in hotels," *J. Travel Tourism Marketing*, vol. 30, nos. 1–2, pp. 23–40, Jan. 2013, doi: 10.1080/10548408.2013.750971.

[42] I. Sutherland, Y. Sim, S. K. Lee, J. Byun, and K. Kiatkawsin, "Topic modeling of online accommodation reviews via latent Dirichlet allocation," *Sustainability*, vol. 12, no. 5, p. 1821, Feb. 2020, doi: 10.3390/su12051821.

[43] D. Li, J. You, K. Funakoshi, and M. Okumura. (2022). *A-TIP: Attribute-Aware Text Infilling via Pre-Trained Language Model*. [Online]. Available: https://aclanthology.org/2022.coling-1.511.pdf

[44] G. Alexander, M. Bahja, and G. F. Butt, "Automating large-scale health care service feedback analysis: Sentiment analysis and topic modeling study," *JMIR Med. Informat.*, vol. 10, no. 4, Apr. 2022, Art. no. e29385, doi: 10.2196/29385.

[45] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[46] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[47] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with ChatGPT," *Mach. Learn. Appl.*, vol. 14, Dec. 2023, Art. no. 100508, doi: 10.1016/j.mlwa.2023.100508.

[48] X. Jiang, Y. Liang, W. Chen, and N. Duan, "XLM-K: Improving cross-lingual language model pre-training with multilingual knowledge," 2021, *arXiv:2109.12573*.

[49] W. Alshehri, N. Al-Twairesh, and A. Alothaim, "Affect analysis in Arabic text: Further pre-training language models for sentiment and emotion," *Appl. Sci.*, vol. 13, no. 9, p. 5609, May 2023, doi: 10.3390/app13095609.

[50] A. Alduailej and A. Alothaim, "AraXLNet: Pre-trained language model for sentiment analysis of Arabic," *J. Big Data*, vol. 9, no. 1, p. 72, May 2022, doi: 10.1186/s40537-022-00625-z.

[51] S. Bird, E. Klein, and E. Loper. *NLTK Documentation*. Accessed: Jan. 26, 2024. [Online]. Available: https://www.media.readthedocs.org

[52] M. Nasseri, P. Brandtner, R. Zimmermann, T. Falatouri, F. Darbanian, and T. Obinwanne, "Applications of large language models (LLMs) in business analytics—Exemplary use cases in data preparation tasks," in *Proc. HCI Int. Late Breaking Papers*, in Lecture Notes in Computer Science, H. Degen, S. Ntoa, and A. Moallem, Eds., Cham, Switzerland: Springer Nature, 2023, pp. 182–198.

[53] K. Dinesh and S. Nathan, "Study and analysis of chat GPT and its impact on different fields of study," *Int. J. Innov. Sci. Res. Technol.*, vol. 8, no. 3, pp. 827–833, Mar. 2023. [Online]. Available: https://zenodo.org/records/7767675

[54] A. Konrad and K. Cai. (2024). *AI Unicorn Anthropic Releases Claude 3, A Model It Claims Can Beat OpenAI's Best*. [Online]. Available: https://www.forbes.com/sites/alexkonrad/2024/03/04/anthropic-releases-claude-3-claims-beat-openai/

[55] *Claude 3: A New Generation of AI*. Accessed: Jun. 24, 2024. [Online]. Available: https://docs.anthropic.com/claude/docs/models-overview

[56] W. Aljedaani, F. Rustam, M. W. Mkaouer, A. Ghallab, V. Rupapara, P. B. Washington, E. Lee, and I. Ashraf, "Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of U.S. airline industry," *Knowl.-Based Syst.*, vol. 255, Nov. 2022, Art. no. 109780, doi: 10.1016/j.knosys.2022.109780.

[57] L. Steven. *Textblob Documentation: Release 0.18.0.post0*. Accessed: Jan. 27, 2024. [Online]. Available: https://readthedocs.org/projects/textblob/downloads/pdf/dev/

[58] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, May 2014, vol. 8, no. 1, pp. 216–225.

[59] U. Naseem, I. Razzak, K. Musial, and M. Imran, "Transformer based deep intelligent contextual embedding for Twitter sentiment analysis," *Future Gener. Comput. Syst.*, vol. 113, pp. 58–69, Dec. 2020, doi: 10.1016/j.future.2020.06.050.

[60] Y. Zhang, "DialoGPT: Large-scale generative pre-training for conversational response generation," 2019, *arXiv:1911.00536*.
[61] OpenAI. *Introducing ChatGPT*. Accessed: May 29, 2023. [Online]. Available: https://openai.com/blog/chatgpt
[62] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, p. 159, Mar. 1977, doi: 10.2307/2529310.
[63] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples," *BMC Med. Res. Methodology*, vol. 13, no. 1, pp. 1–7, Dec. 2013, doi: 10.1186/1471-2288-13-61.
[64] M. Vijaymeena and K. Kavitha, "A survey on similarity measures in text mining," *Mach. Learn. Applications: Int. J.*, vol. 3, no. 1, pp. 19–28, Mar. 2016, doi: 10.5121/mlaij.2016.3103.
[65] R. Han, T. Peng, C. Yang, B. Wang, L. Liu, and X. Wan, "Is information extraction solved by ChatGPT? An analysis of performance, evaluation criteria, robustness and errors," 2023, *arXiv:2305.14450*.
[66] P. Rostami, A. Salemi, and M. J. Dousti, "PersianMind: A cross-lingual persian-english large language model," 2024, *arXiv:2401.06466*.
[67] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, and W. Han, "ChatIE: Zero-shot information extraction via chatting with ChatGPT," 2023, *arXiv:2302.10205*.

TAHA FALATOURI received the B.A. degree in industrial management from Allameh-Tabataba'i University and the master's degree from the Tarbiat Modares University of Tehran. He is currently pursuing the Ph.D. degree with Thomas Bata University in Zlín, Czech Republic, specializing in industrial engineering. With a rich background spanning 12 years in banking, manufacturing, and retail management, he has been focusing on his academic journey, since 2019. His research has centered on analytics and operational research, particularly in customer experience analysis. He employs cutting-edge AI and machine learning techniques to drive innovation and address complex challenges in enhancing customer satisfaction.

DENISA HRUŠECKÁ received the master's degree in industrial engineering, in 2009, and the Ph.D. degree in production planning and scheduling, in 2015. She is currently a Senior Lecturer and a Researcher with the Faculty of Management and Economics, Tomas Bata University in Zlín. Her current research interests include advanced production planning and scheduling methods, including logistics, sales activities, and customer behavior monitoring. She also deals with process management and optimization in all business areas, focusing on the new trends of artificial intelligence.

THOMAS FISCHER received the B.A. degree in electronic business from the University of Applied Sciences Upper Austria, Steyr, in 2012, and the joint M.Sc. degree in digital business from the University of Applied Sciences Upper Austria and the Johannes Kepler University Linz, in 2014. He is currently pursuing the Ph.D. degree in social sciences, economics and business with Johannes Kepler University Linz.

He was a Researcher with the University of Applied Sciences Upper Austria, Johannes Kepler University Linz, and the University of Passau. He is also a Research Associate with the University of Applied Sciences, Steyr, Austria, specializing in information systems, digital transformation, and artificial intelligence. He was also a Visiting Researcher with Missouri University of Science and Technology, in 2014, and the University of South Florida, in 2018. He has published more than 40 articles so far and is serving as a reviewer for various journals on a regular basis. He is also a Founding Member of the NeuroIS Society and serves as an Associate Editor for the International Conference on Wirtschaftsinformatik.