IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2024, accepted 1 August 2024, date of publication 9 August 2024, date of current version 22 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441514

SURVEY

A Systematic Review on Responsible Multimodal Sentiment Analysis in Marketing Applications

INÊS CÉSAR[1], IVO PEREIRA[1,2,3,4], FÁTIMA RODRIGUES[1,2], VERA L. MIGUÉIS[2,5], SUSANA NICOLA[1,2], ANA MADUREIRA[1,6], (Senior Member, IEEE), JOSÉ LUÍS REIS[7,8], JOSÉ PAULO MARQUES DOS SANTOS[7,8], AND DANIEL ALVES DE OLIVEIRA[4]

[1]Interdisciplinary Studies Research Center (ISRC), Instituto Superior de Engenharia do Porto, 4249-015 Porto, Portugal
[2]Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC), Faculdade de Engenharia da Universidade do Porto, 4200-465 Porto, Portugal
[3]Faculdade de Ciência e Tecnologia, Universidade Fernando Pessoa, 4249-004 Porto, Portugal
[4]E-goi, 4450-190 Matosinhos, Portugal
[5]Faculdade de Engenharia da Universidade do Porto, 4200-465 Porto, Portugal
[6]INESC INOV-Lab, 1000-029 Lisboa, Portugal
[7]Department of Business Administration, Universidade da Maia, 4475-690 Maia, Portugal
[8]LIACC—Artificial Intelligence and Computer Science Laboratory, University of Porto, 4200-465 Porto, Portugal

Corresponding author: Ivo Pereira (iaspe@isep.ipp.pt)

This work was supported by the National Funds through the FCT—Fundação para a Ciência e Tecnologia within the PHYNHANCAI Project (http://doi.org/10.54499/2022.01303.PTDC).

ABSTRACT The intrinsic challenges of contemporary marketing encourage discovering new approaches to engage and retain customers effectively. As the main channels of interactions between customers and brands pivot between the physical and the digital world, analyzing the outcome behavioral patterns must be achieved dynamically with the stimulus performed in both poles. This systematic review investigates the collaborative impact of adopting multidisciplinary fields of Affective Computing to evaluate current marketing strategies, upholding the process of using multimodal information from consumers to perform and integrate Sentiment Analysis tasks. The adjusted representation of modalities such as textual, visual, audio, or even psychological indicators enables prospecting a more precise assessment of the advantages and disadvantages of the proposed technique, glimpsing future applications of Multimodal Artificial Intelligence in Marketing. Embracing the Preferred Reporting Items for Systematic Reviews and Meta-Analysis as the research method to be applied, this article warrants a rigorous and sequential identification and interpretation of the synergies between the latest studies about affective computing and marketing. Furthermore, the robustness of the procedure is deepened in knowledge-gathering concerning the current state of Affective Computing in the Marketing area, their technical practices, ethical and legal considerations, and the potential upcoming applications, anticipating insights for the ongoing work of marketers and researchers.

INDEX TERMS Affective computing, customer behavior, marketing, multimodal artificial intelligence, sentiment analysis, systematic review, trustworthy AI.

I. INTRODUCTION

The constant evolution of Digital Marketing has led to an inevitable search for new techniques to improve strategies and maximize the brand's success with the intended target audience [1], [2]. Understanding and interpreting human interaction, between the physical and digital worlds, takes into account various factors that cut across Marketing domains, requiring the choice of tools to make the process flexible to the inherent context [3], [4], [5]. The use of Artificial Intelligence (AI) offers a wide range of alternatives [6], highlighting consumer patterns as the basic knowledge for brands to build successful campaigns customized by their preferences [7], [8], [9]. Marketers are already taking as current work tools the developed systems due to the simplicity of using the implementation such complex models, which

The associate editor coordinating the review of this manuscript and approving it for publication was Okyay Kaynak.

can leverage dependencies and patterns not captured by humans [10]. The effectiveness of the provided applications is one of the topics currently under investigation [2], [4], [5], [8], [10], with the inclusion of emotional cues on the data under study being one of the most explored [2], [7], [9], [11], [12], [13]. As stated by Cesar et al., Gandhi et al, and Tomar et. al [12], [14], [15], the extraction and application of multimodal instead of unimodal data resources, considered to represent sensory modalities expressed or perceived and with heterogeneous qualities, arises the possibility to balance the complexity of the performance and apply heterogeneous complementary information. By guaranteeing a data structure preserves the existing relation of different data modalities, consumers' preferences can generate contextual knowledge for brands [7], [12] and scrutinize emotions and feelings of their opinions to improve their service [6], [9], [12]. Expressive forms of communication differ considerably from physical to digital reality, with different approaches of affective computing tasks [4]. Research about the consequences of the simulated stimuli of human emotions covers the entire journey of customer engagement, capturing the latest preferences and how to predict and incorporate them into the next marketing strategy [5], [10], [13]. Prediction, attention-based, and classification mechanisms are some algorithms employed to cover the analysis of verbal and non-verbal data [11], [13].

Affective Computing is a multidisciplinary area that seeks to bring together the benefits of research carried out on coding systems that can commune, trigger or speculate human emotions [16], [17], [18], [19], [20], [21]. Technology-enhanced systems guarantee a place of great importance in other fields of business, demonstrating the flexibility and importance of their action in understanding and personalizing the user experience [17], [21]. Understanding human emotions is a complex process [18], [19], marking the evolution of bringing technology closer to consumers' cognitive abilities, allowing these systems to respond more effectively to their needs [20]. The innovative and recent progress of consulting Affective Computing approaches to learn through human interactions has generated several scientific inputs about the matter, enhanced by the capture and study of multimodality inherent in shoppers behavioral and psychological signals [17], [18], [19]. However, as the use of sentiment analysis and emotion recognition are becoming widespread tasks employed in various areas [17], [20], [21], measures need to be taken to ensure that these systems are resourceful in reliable environments and protected from any improper use or manipulation of the data input [22], [23], [24], [25], [26]. The reinforcement of ethical and regulatory considerations has gained increased attention worldwide, with several governmental organizations in the race for legislative transformation on these technological tools [22], [25], [26]. The mass adoption of AI made nations, like the European Union, take further steps to create a draft of a worldwide regulative procedure on AI, to responsibly measure and prevent points of failure in users' privacy, protection, and authentication [22], [23], [24], [25], [26]. To fulfill the requirements of a reliable application, developers and creators of intelligent content are called upon to work to elevate the stability and security [22], [23], giving a sense of trust and confidentiality in AI to the customer [24], [25], [26].

This study seeks to analyze the current state of multimodal approaches in investigating consumer sentiment across various social interactions, aiming to enhance marketing strategies through the integration of Affective Computing and Multimodal Artificial Intelligence (MAI) [27]. By conducting a systematic review of the current advancements in MAI, the research reflects on scientific progress in the field and investigates the potential combination of sentiment analysis and emotion recognition across physical and digital contexts. The proposed approach provides an in-depth look into implementing AI models capable of recognizing emotions from diverse customer modalities, ultimately aiming to improve the efficacy of marketing in a phygital world. The project addresses the current state of marketing developments and the knowledge base required for integrating different AI modalities, while also emphasizing the importance of ethical and legal considerations in distinguishing between low and high-risk practices. Utilizing the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) method, as highlighted by Sohrabi et al. [28], ensures transparency and reproducibility in the systematic review process, which is crucial for establishing a reliable foundation in multimodal sentiment analysis. The complexity of integrating various modalities necessitates a rigorous and standardized approach to review and validate the results systematically, distinguishing this work from similar studies by Prabowo et al., Gandhi et al., and Tomar et al. [11], [12], [15], which lack comprehensive application of PRISMA protocols. This research aims to provide insights and practical implications for future marketing strategies, addressing the evolving landscape of consumer engagement and retention.

The article follows this structure: in Section II, the systematic review method outlines all the steps conducted for the research. Section III includes the details of the accomplished results, presents a discussion on the related work, and highlights both technical and practical issues of the research questions. Finally, in Section IV, the conclusions are derived, and associated with a list of all the outlined constraints, both remembered and forgotten by the included authors, along with their limitations and future directions.

II. METHOD

The conducted systematic review employs the PRISMA method to accurately consolidate the most recent scientific contributions to multimodal Marketing on customer sentiment. This systematic review stands out from the previously analyzed [11], [12], [15] by using this approach to seek answers to the application of sentiment analysis and how dynamically is making up the work of marketers in their strategies. The contextualization of the research relies on demonstrate a wider view on the modalities covered by enterprises and the projected integration of information, both shown or expressed by the consumer, with trustworthy emotional cues. As so, the responsibility to highlight ethical and regulatory considerations is crucial, detailing both challenges and opportunities provided with the merge of knowledge bases. The future directions are also propelled by carefully leveraging all the included articles, covering with an hybrid appreciation in each application.

This methodology is based on a series of procedures to be followed so that all the knowledge returned and documented can be replicated and analyzed to compare the conclusions drawn. As described in [28], PRISMA highlights the need to formulate research questions that contain the main points about the domains in question. To answer them correctly, the identification of the main areas of research contextualizes the articles returned by creating a computational instruction that translates the relationship between the context, the population, and the associated concept. The information, obtained from specific repositories, completes the initial base of publications to be submitted to a selection process. Nevertheless, only through the development and consideration of inclusion and exclusion criteria is it possible to positively or negatively validate the article and consider it in the review. Adjusting the practice of PRISMA instructions guarantees continuous filtering until the best resources are obtained.

A. RESEARCH QUESTIONS

The exploration of different alternatives for implementing multimodality with Affective Computing is suggested for enhancing new and uncovered Marketing strategies, connecting digital and physical realities. A series of research questions are formulated as guidelines to broadly understand the symbiotic relation between Marketing and Affective domains.

TABLE 1. Research questions.

Identifier	Research Question
RQ1	What is the current state of research in Multimodal Sentiment Analysis for Marketing?
RQ2	How do Multimodal Sentiment Analysis practices perform across different Marketing domains?
RQ3	What are the key modalities and datasets commonly used for Multimodal Sentiment Analysis?
RQ4	How is achieved the integration of multimodal data with technical models for Multimodal Sentiment Analysis?
RQ5	What ethical and regulation considerations are associated with the use of Multimodal Sentiment Analysis in Marketing?
RQ6	What are the challenges and opportunities for potential future applications of Multimodal Sentiment Analysis in Marketing?

First, an inquiry into the current state of Multimodal Sentiment Analysis for Marketing as a basis operation to apprehend the existing expansions on using emotional cues to analyze and predict consumer behaviors and supplement the next interactions with insights from the obtained conclusions. Then, an explanation of the beneficial results performed by Multimodal Sentiment Analysis in Marketing domains is provided to evaluate the evolution of some of the most popular strategic chores. After recognizing the dynamics of communication channels between stakeholders of Marketing approaches, the research underlines the different modalities that personalize the interactions and are protagonists in the sentiment annotated data collections. In addition to this, a detailed exploitation of techniques and methods used to integrate each feature into the learning progress of AI models based on sentiment analysis is conducted regarding each combination of variables for transversal comprehension. The following debrief regards the analysis of ethical and legal restrictions currently imposed to monitor the good use of these frameworks for research or commercial use. Finally, an investigation of the challenges and opportunities faced for designing and innovating sentiment-based AI tools with contextual Marketing data. Table 1 identifies the sequence of research questions that will be answered in the results of retrieved publications section.

B. SCIENTIFIC REPOSITORIES

The selection of scientific repositories represents one of the first decisions to conduct a systematic review. As noticed in the work of [28], the 2020 update on the PRISMA method requires, in the first phase named *Identification*, the declaration of each data source from where the records were collected to enhance transparency throughout each data source research task. Web of Science, IEEEXplore, and Science Direct were the three data repositories chosen for this systematic review, as detailed in Table 2. The parameters in each database are defined to mitigate major differences between them and will be explored later in this section.

TABLE 2. Scientific repositories.

Identifier	Repository	URL
SR1	Web Of Science	https://www.webofscience.com/
SR2	IEEEXplore	https://www.ieeexplore.ieee.org/
SR3	Science Direct	https://www.sciencedirect.com/

C. SEARCH TERMS

The definition of search terms is carried out to give contextual linkage among the research fields and the planned investigation objectives. Providing a combination of keywords that correlate with the specified domains orientates the retrieval of articles, offering a more reliable spectrum of scientific contributions [28].

TABLE 3. Search terms.

Domain	Keywords
Multimodal AI	"Multimodal"
Affective Computing	("Sentiment Analysis" OR "Emotion Recognition")
Marketing	("Client" OR "Customer" OR "Consumer")

As presented in Table 3, three domains were identified as being the major areas to explore. The application of terms such as *Multimodal* or *Affective Computing* scopes specifications to the AI applications for being an area with a large diversity of applications. Nevertheless, the consideration for the domain of Marketing was to incorporate the experiences established with the target audience, regarding its analysis. Due to that, a combination of keywords for Affective Computing and Marketing were expressed so that the concatenation of all would be included in a single string query, as portrayed in Table 4.

TABLE 4. Research query.

```
"Multimodal" AND ("Sentiment Analysis" OR "Emotion Recognition")
AND ("Client" OR "Customer" OR "Consumer")
```

D. INCLUSION AND EXCLUSION REQUIREMENTS

Inclusion and Exclusion requirements play one of the most influential roles in the selection of the final set of studies to be included in the review. Throughout the entire research scheme, the defined restrictions speed up the decision-making responsibility for each article screening. Along with that, some of these parameters are even included as an advanced search rule for the attained data. Tables 5 and 6 encode via an identifier all the requirements incorporated to justify the verdict of some articles not meeting and others nearly missing out on the needful contribution to develop a systematic review, being one of the new changes made to the PRISMA checklist and documented in [28].

TABLE 5. Inclusion requirements.

Identifier	Inclusion Requirement
IR1	The article is part of a collection of peer-reviewed publications
IR2	The article belongs to the field of Computer Science, NeuroMarketing and Emotional Computation
IR3	The article is focused on contributing with relevance to the study domains
IR4	The article describes a system, a framework, or an application scenario with both theoretical and practical knowledge bases
IR5	The article has evidence of the evaluation and validation of the proposed conclusions

TABLE 6. Exclusion requirements.

Identifier	Exclusion Requirement
ER1	The article is over 3 years
ER2	The article is not written in English
ER3	The article is not from a journal or a conference proceeding
ER4	The article is not adaptable to components from included and related domains
ER5	The article is focused only the application of AI on unimodal approaches
ER6	The article is either focused on using AI to recognize other psychological traits, such as sarcasm or any other emotion individually, or to recognize other information not related to emotional features

E. PUBLICATIONS EXTRACTION

The extraction of publications took place under the assumptions of the PRISMA method, defined by three phases: *Identification*, *Screening*, and *Included*. Figure 1 shows the diagram updated to the 2020 version, which defines the process of evaluating all the articles through a series of inferences about their composition before they are included.

In the *Identification* phase, the number of articles collected from each specified repository was highlighted. The advanced search criteria also defined by SR1, SR2, and SR3 refer to the range of publication years (between the first day of 2020 and the last day of 2023), written in English and belonging to peer-reviewed formats, specifying journal and conference articles. This data guarantees compliance with the inclusion requirements, identified as IR1, IR2, and IR3 respectively in the Table 5 from the previous section. However, specific commands referent to the advanced search in each repository could not be avoided, applying different and unique filters for each virtual database. For SR1, the definition of command filter ''ALL'' extended to all the query terms, customizing an wider access on articles. As a result, a total of 31 articles were retrieved. In SR2, the ''Full Text and Metadata'' command was set to interpret the query as the most complete, similar to the command used in SR1. In addition to the presence of the open access option, the list of results was organized by the relevance of each article. Since the total result was more than 10,000 references with the previous details configurated, the first 500 references with more relevance were selected. In SR3, the query was introduced without the need to specify any filtering commands to the search, presenting a selecting option of the type of articles only with the description ''research articles'', returning 490 results. The use of tools such as Mendeley helped to identify duplicate articles, with only 2 being found in this case. Since no automation tool was used or a reason was found to exclude any more articles, the identification phase resulted in the capture of 1019 articles.

The *Screening* phase corresponds to the corpus of the entire review and consists of three periods of reviewing the articles, increasing the detail in the processing from the first to the last. The first phase involves scanning all the abstracts and benchmarking the selection concerning the objectives of this review: the relevance of the article, the main topics discussed are accordant to the research questions formulated, the match with one or several inclusion requirements, and the precisive search for a match with any exclusion ones. This process led to the exclusion of 661 articles, reducing the number of articles that went through the re-evaluation stage to 358. Based on the information in the abstracts, the reading of information in the introduction was introduced as a complement to this phase, where it was possible to consult the objectives of each article and find proposals that met at least one of the inclusion criteria. Examples of articles such as systematic reviews, other types of reviews, or even surveys were some of the reasons for excluding 208 publications. Finally, a full reading of all the articles made it possible to identify those that met both the inclusion and exclusion criteria. The specific reasons can be summarized in ER4,

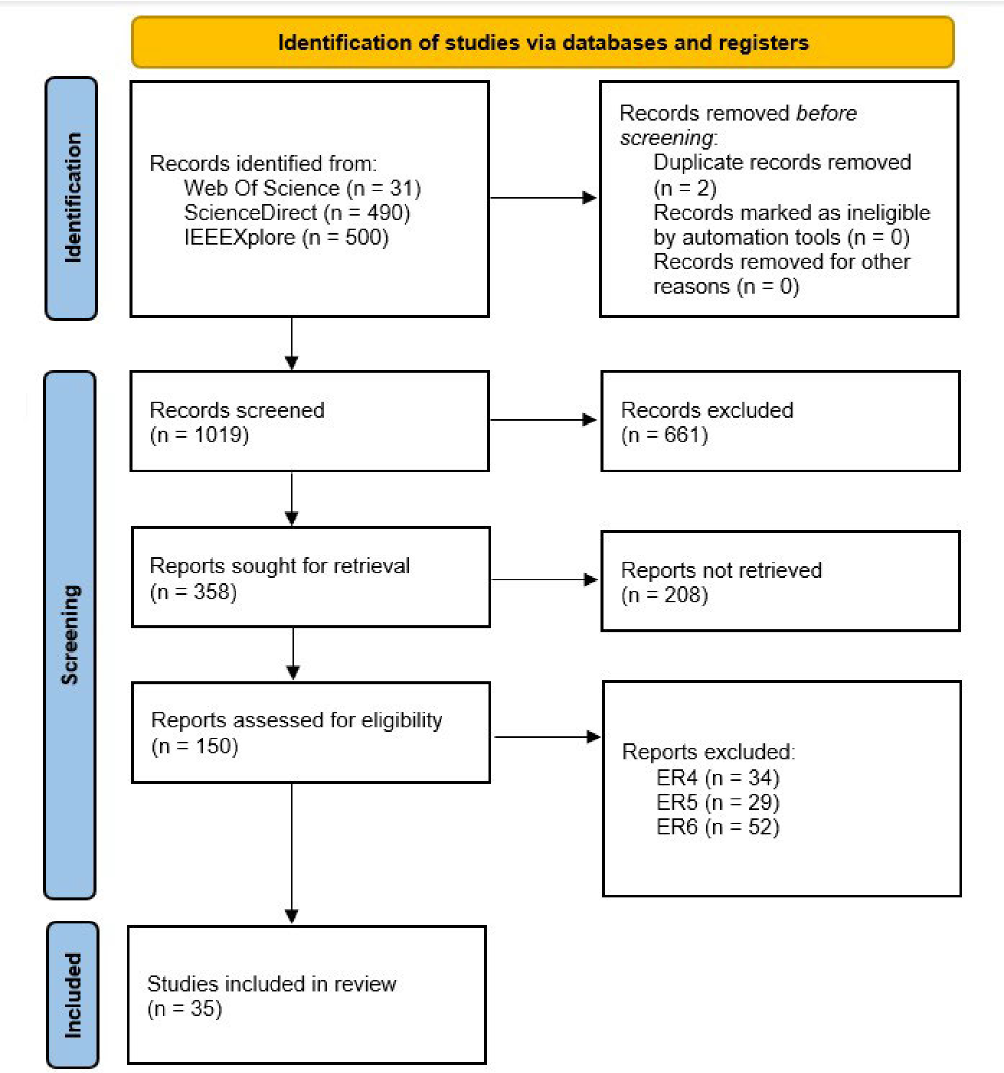
Identification of studies via databases and registers
Identification
Records identified from:
Web Of Science (n = 31)
ScienceDirect (n = 490)
IEEEXplore (n = 500)
Records removed *before screening*:
Duplicate records removed (n = 2)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Screening
Records screened (n = 1019)
Records excluded (n = 661)
Reports sought for retrieval (n = 358)
Reports not retrieved (n = 208)
Reports assessed for eligibility (n = 150)
Reports excluded:
ER4 (n = 34)
ER5 (n = 29)
ER6 (n = 52)
Included
Studies included in review (n = 35)

FIGURE 1. PRISMA Diagram with the retrieved articles, according to [28].

ER5, and ER6, with 34, 29, and 52 articles disregarded, respectively.

Finally, in the *Included* phase, the number of articles included in the search for answers to the defined research questions is defined. In this specific case, a total of 35 articles were selected. Figure 2 offers an illustrated distribution per year and per scientific repository of the final result of publications considered. The most referenced year is 2022, followed by 2023, expressing the new-found applications that populate this systematic review environment.

III. RESULTS AND DISCUSSION

The following section details the general attributes of the obtained articles with the application of the PRISMA methodology. It aims to document, for each research question, the number, the main topics and the pertinence of their selection to reply properly to the investigation inquiries.

A. MULTIMODAL SENTIMENT ANALYSIS IN MARKETING: CURRENT STATE OF RESEARCH

Marketing efforts currently face challenges and adversities due to fluctuations in customer behavioral trends [29], [30], [31], [32]. With the increasingly active parallel presence of consumers in the digital world [30], [33], [34], [35], [36], [37], it is necessary to foresee how complementarity with the real world can benefit brands [38], [39], [40], [41], [42], [43]. The multimodal approach is therefore at the forefront of the most knowledge-intensive alternatives [31], [39], [40], [41], [44], [45], [46], [47].

For the first research question, the conducted search returned 30 articles with relevant information that allows us to describe the current state of Marketing from a multimodal perspective [32], [33], [48], [49], [50], [51], with different categories of scientific contributions. In addition to practical applications that contextualize this innovation

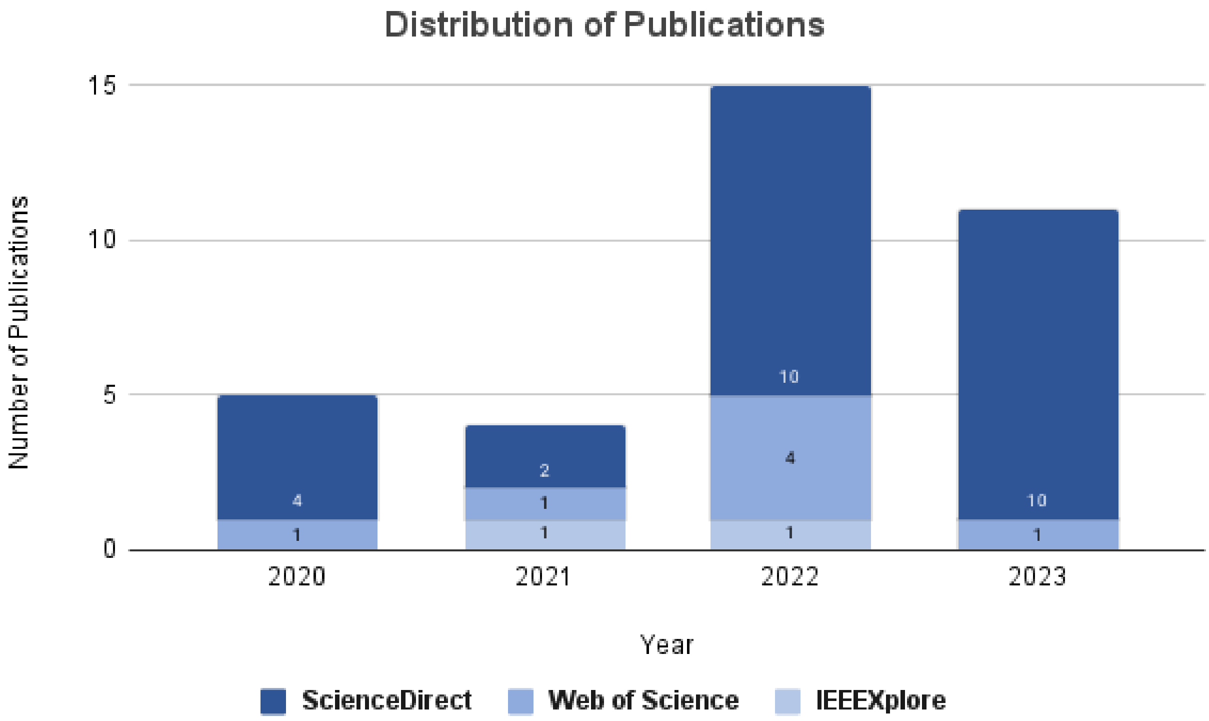

FIGURE 2. Distribution of Retrieved Publications by research year and by repository.

in the real world, to be reviewed and analyzed in more detail later, articles were also obtained considering the prosperity of new forms of communication through the use of AI [29], [30]. Taking advantage of current technology, and combining it with the capture and transmission of information in different formats [31], [38], [39], [40], [41], [44], [46], [47], [50], makes it possible to monitor consumer perceptions of different stimuli produced and simulated by companies. In this way, multimodality makes it possible to aggregate all the knowledge with a common purpose and ensure that a relationship is established between all the heterogeneous evidence produced [29], [30], [35], [43], [45]. The emotional weighting of these interactions linked to the multiple interpretations of data raises the uniqueness of new options for marketers [37], [38], [39], [40], [41], [49].

The inclusion of Affective Computing has recently become notable as one of the ways to target customer preferences to positively influence the development of subsequent actions. It is based on these considerations that [29] conceives a framework capable of establishing collaboration between AI and Human Intelligence in Marketing, focusing on possible system optimizations. This relationship creates a symbiotic environment, bringing together the strengths of both intelligences and co-operate on solving issues related analytical efforts and emotional robustness. Grewal et al. [52] reveals that the dynamics of the elements throughout the consumer's physical and digital journey is a crucial factor, emphasizing knowledge that is more enriched by the different modalities. Identifying the different forms of communication between stakeholders, which influence the success of Marketing actions, is carried out to anticipate and expand the research of retailers and researchers.

B. MULTIMODAL SENTIMENT ANALYSIS PERFORMANCE ACROSS MARKETING DOMAINS

After gaining an in-depth knowledge of the present status of Multimodal Sentiment Analysis in Marketing, the next step is to figure out how the different Marketing departments benefit from this method. Evidence of practical applications in the various fields of applied strategies can be found in 28 articles. Some of the most noteworthy were applications aimed at analyzing social networks [30], [31], [34], [35], [36], [37], [43], [49] reviewing products [31], [32], [45], [53] and also predicting elements such as the prices of services [54] and the popularity given by consumers [33], [48], [51]. The importance of studying direct and indirect interactions between companies and customers, evaluating the dialogue established [44], [47], [50] and the public opinion generated is also highlighted [39], [40], [41], [46], [55], [56], [57], [58]. Even if a report is categorized as follows, for a better grasp of the information on the issue, it is worth highlighting the dependence between these domains and their effectiveness.

Nowadays, an assiduous presence on social networks is becoming a general responsibility for most organizations due to the ease of instant delivery of advertising [30], [34], [35], [36], [42], [43]. This kind of communication, which is widely used several times a day for long periods by society in general, warrants special attention due to its convenience as an e-commerce platform. This increases customer engagement for more information about the various services on offer and their purchase [30], [34], [35], [36], [37], [42]. Although it is referred to as an area of Marketing application under investigation [30], [34], [35], [36], [37], [42], [43], multimodality has a positive influence on attracting new consumers and consolidating existing ones. The customization of these exchanges is dependent on the context of the business sphere [30], [34], [36], [43], but ensures more flexible participation by the parties involved, allowing them to create content that can be integrated into subsequent strategy updates [34], [35], [42]. As a result, quickly obtaining opinions by associating emotional expressions [34], [35], [42] gives way to understanding current customer trends [30], [34], [35], [36], [42] by analyzing the evolution of likes, comments, and shares using likehood models [30], [34], [36], [43].

The employment of the multimodal approach in service or product review analysis has a positive effect on consumer engagement [31], [32], [45], [46], leading to an increase in e-commerce sales [31], [32], [46]. This is what Mehbodniya et al. [32] explores, where they assess product preference in an online context by applying sentiment analysis to product reviews. Alongside this, they also highlight how multimodality provides more information about the product [31], [45], [46], reducing customer uncertainty at the time of purchase [31], [32], [46]. However, there are problems in both sides associated with the misleading multimodal information, such as the increase in product returns bought online for not matching the expected product sent by brands [31], [32] and the discrepancy in the obtained data reviews from questionnaires aimed at regular customers [31], [45], [46]. The heterogeneity of the characteristics obtained by real product evaluation [31], [32], [45] interconnects the both problems to a single solution, focused on data fusion as a way of overcoming these situations [31], [32], [45], [46], highlighting the research work considered by the authors [31], [45] to be preliminary on these techniques. They see the use of emotion stimulated in these reviews, some through tone and frequency of voice [31], [46], as a way of accurately predicting future audience behavior [32], [45].

It is also a common procedure to use multimodality to grant improvement on structure recommendation systems covering more irregular use cases, simulating the complexity of real ones [33], [49], [54]. Combining the diversity of variables, ensuring sensory samples of the same information ensures better alternatives of developing advertising [33], [48], [54], recognizing not only the content as the basis of all the dynamics but also the foreseen target audience [33], [48], [51]. The use of social networks, capable of supporting the recognition and development of new content [48], [49], [54], allowing it to extend its prediction to other outcomes but made up of the union of multimodal representations [33], [48], [49], [51], [54]. The performance of predictive models expects the flexibility of applying them to a phygital reality [51], [54].

The use of sentiment analysis in dialog interactions [44], [47], [50] provides a new angle of perception that is difficult to capture in modality with text in images. Developments in interpreting conversations with feelings [44], [47], [50] are still a growing topic of study and practice due to the associated complexity [44], [50]. This type of data structure validates interactions between humans as well as robots or other types of responsive systems [50], making it plausible to combine with other technologies. Both this and the category referring to opinion reviews carried out voluntarily by consumers name emotions as a viable way of evaluating their interactions [39], [40], [41], [46], [55], [56], [57], [58]. Social networks once again play a leading role as a platform where opinions are obtained in different formats [39], [40], [41], [46], [55], [56] and also capture non-verbal data in addition to verbal data. Multimodality is once again chosen as a way of involving the emotions associated with opinions to capture them in more detail [39], [40], [41], [46], [55], [56], [57], [58].

C. KEY MODALITIES AND MULTIMODAL DATASETS FOR SENTIMENT ANALYSIS

The research carried out returned 30 articles that allow for the analysis of applications and technical developments that exemplify a multimodal approach to the tasks and challenges of the various Marketing strategies [30], [31], [32], [33], [34], [35], [36], [37], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [53], [54], [55], [56], [57], [58], [59], [60]. These, in turn, extend to sentiment analysis which, through the richness represented in the heterogeneity of the data, guarantees a better understanding of current dynamics. The exponential growth of data captured, both verbal and non-verbal, motivates the adoption of multimodal approaches that allow an investigation centered on customer preferences.

By analyzing the content documented by the different articles, we can see that 5 categories reflect the combination of modalities, expressed in Figure 3. The circular graph confirms the dominance of the fusion between textual and visual features present in 47% of the articles [30], [32], [33], [34], [35], [36], [37], [38], [41], [43], [44], [48], [55], ideally justified by the fact that these are two of the most common forms of information sharing in phygital Marketing. However, the addition of audio to these is proving to be an increasingly interesting alternative consisting of knowledge structures with evidence of immediate contact [31], [39], [40], [46], [47], [49], [50], [51], [56], [57], [58], [60]. In third place, with 7% of uses, audio evidence is replaced by audio and visual resources, guaranteeing greater diversity accompanied by high complexity [45], [59]. The succession of modalities that can be broken down or represented in other formats reveals the challenge of granularity and how this increases the difficulty of the system's performance. Even so, the disparate foundations of information provided

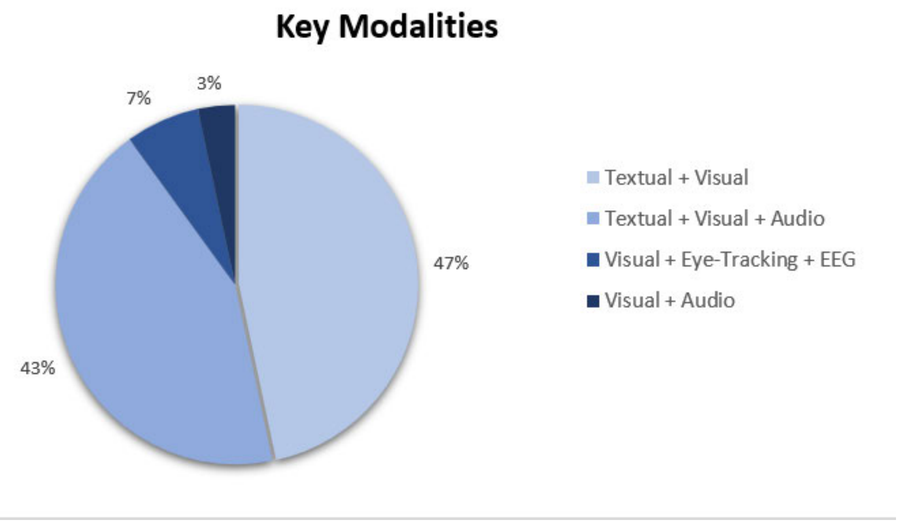

FIGURE 3. Modalities used in retrieved articles.

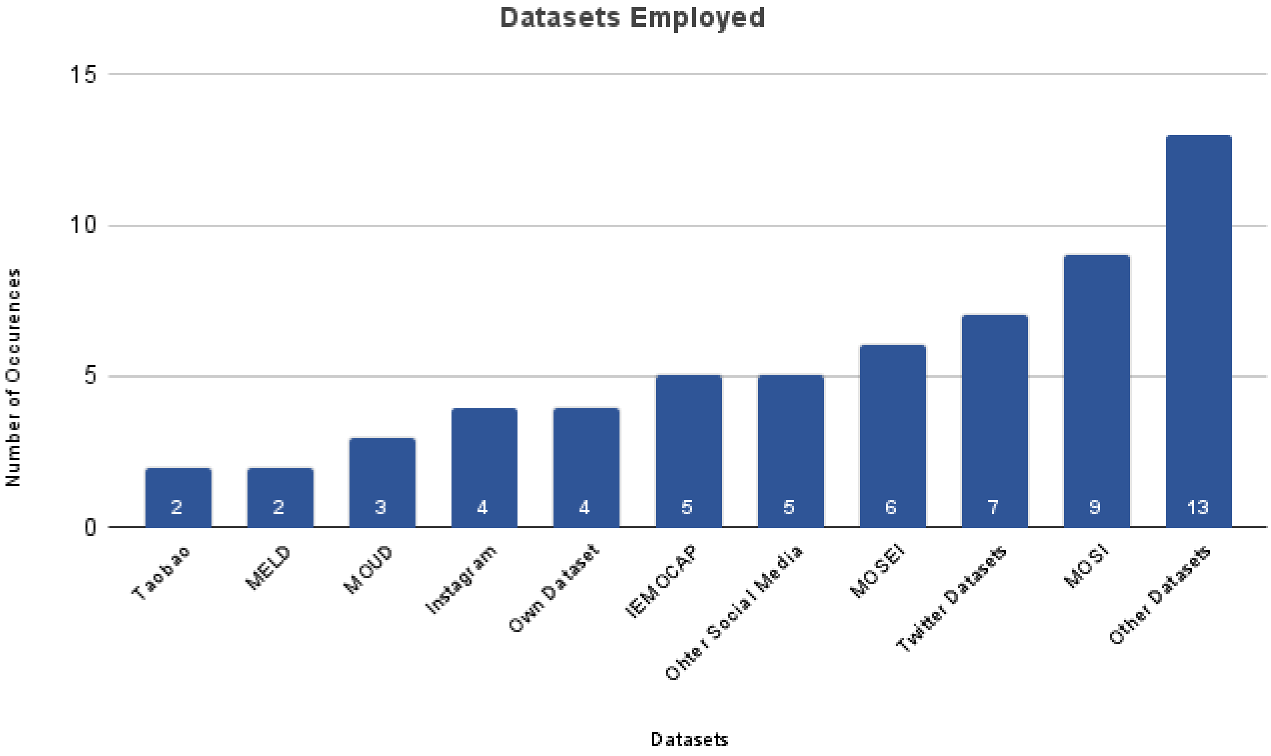

FIGURE 4. Datasets employed in retrieved articles.

by the model highlight the integral progress in generating knowledge. Other categories, such as the inclusion of series capturing Eye-Tracking and Electroencephalography (EEG), complementary with visual variables, or even the study of the impact of audio only with visual or with textual modalities, are the least mentioned, but they are novel because of their computation with AI. In this way, and in line with the inequalities of the variables included, their integration is unique and therefore relevant to be investigated.

The process of integrating the different modalities in multimodal sentiment analysis tasks raises a series of challenges to consider to validate the knowledge generated. Amongst these, it is possible to consider that the choice of data set generates the first of the restrictions affecting the results obtained. The virtue of using data that has already been processed rather than data in its natural state is one of the best practices to be carried out to mitigate any existing discrepancies. However, the requirement for multimodality in the information used makes it more complex to administer techniques that model the data correctly. Figure 4 graphically presents the datasets the scientific community adopts on emotional categorization in human interaction activities in Marketing. These datasets are great resources for performing studies and experiments about the relation of different modalities, combined with emotional labels for classification tasks or providing numerical values that measure each emotion's valence. Datasets such as Taobao [31], [48], CMU-MOSI [40], [46], [47], [49], [51], [53], [56], [57], [58], CMU-MOSEI [40], [46], [49], [51], [57], [58], and MOUD [39], [53], [56] aggregate data samples referring to reviews and feedback expressed by consumers on various products, followed by features that translate the valence or feelings felt employing numerical or categorical variables, depending on the case. Datasets such as MELD [44], [50], IEMOCAP [39], [44], [49], [50], [60], and "Other Datasets" [32], [34], [35], [42], [43], [46], [48], [54] incorporate emotion recognition by keeping data on dialogues, video communications, or even sensory expressions captured by more specific equipment. Datasets made up of inferences from various social networks [32], [33], [35], [38], [42], [43], [53], such as the speed of Instagram [30], [35], [36], [43] and Twitter [33], [34], [35], [37], [55]. support the magnitude associated with the study of normally informal and reliable opinions on the satisfaction of the target audience. However, there are still some cases [41], [45], [49], [59] in which the authors themselves advocate the personalized construction of their datasets using the concatenation of pre-existing data with information specific to the problem itself. The diversity of the aforementioned datasets structures the need to conclude the intermediate and final effects of their use, even if they all positively contribute to the affectivity balance. Regardless of the dissonances between the annotations of each data record, these are even more pronounced with the idealized configuration in each system.

D. TECHNICAL INTEGRATION OF MULTIMODAL DATA

The analysis of modality diversity carried out previously motivates the extension of the research to obtain a better understanding of the different practical alternatives for data integration. Following the same categories shown in Figure 3, this research question takes 29 articles with technical practices of extracting and preprocessing data, as well as the methods adopted for multimodal fusion and classification tasks to retrieve the underlying emotional content. Through evidence expressed in textual, visual, both images and videos, and audio samples samples, it is possible to perceive the associated valence [30], [31], [32], [33], [34], [35], [37], [40], [41], [42], [43], [45], [47], [48], [51], [53], [54], [55], [56], [57], [60], the human emotion expressed in feelings [46], [49], [59], or both [36], [39], [44], [50], [58]. Tables 7, 8 and 9 present a resume of each technique specified for integrating the different modalities in each article.

1) TEXT AND IMAGE INTEGRATION ANALYSIS

Based on the review undertaken, it is possible to identify text and image as the two most commonly used features for building a multimodal perspective in AI integration. Even though not all of them predict sentiment using emotional AI models, as is the case in the articles [30], [42] where polarity is calculated using the number of likes and comments to test hypotheses about the impact of multimodality, it is possible to find similarities in the remaining procedures. In the extraction and pre-processing of textual data, tools such as Glove [37], [41], [43], [44] for word embeddings, Google Cloud Vision API [30], [33], [35], [42] both for textual extraction in its original modality or through OCR on images, and even customized versions of the Bidirectional Encoder Representations (BERT) pre-trained model [34], [55] are commonly used. Other techniques have also been mentioned that complement the success of this task: the use of Linguistic Inquiry and Word Count (LIWC) software [30], Fourier-Bessel Transform (FBT) filter which extends the work of the Latent Dirichlet Allocation (LDA) method [48], the adoption of neural networks such as Convolutional Neural Network (CNN) [44], Bidirectional Long Short Term Memory (Bi-LSTM) for contextual features [41], the addition of attention mechanisms [55] or Bidirectional Gated Recurrent Unit (Bi-GRU) [37] for semantic details and the combination of Latent Semantic Analysis (LSA) and Term Frequency-Inverse Document Frequency (TF-IDF) algorithms [32]. Along side with these, the use of libraries such as Natural Language Toolkit (NLTK) [33], [42], [43], String and res [42], Selenium and Beautiful Soup [36] must be highlighted. There is also the integration of text obtained from image captions using techniques like Computer Graphics Metafiles (CGM) [55] or Bag of Visual Words (BoVW) [35]. In addition to its use for textual extraction, the Google API mentioned above is also used to extract visual features [30], [33], [35], [42], competing with tools such as Visual Geometry Group Networks (VGGNet) [48], [55], Visual Transformers (ViT) [34], CNN [43], [44], LSTM [48], Residual Networks (ResNet), and may [43] or may not [41] have an associated attention mechanism. Similarly, others can be applied, such as scraper software [36] and methods such as Speeded Up Robust Features (SURF) and Multi-Scale Local Binary Pattern (MLBP) [32]. The diversity of options for treating each modality makes it possible to obtain different features contextualized to the problem in question, making

TABLE 7. Summary of integrations with textual and visual modalities.

Ref.	Textual Processing	Visual Processing	Fusion Approach	Models	Baselines
[30]	LIWC, GVAPI	GVAPI	Hybrid	Likelihood	Poisson Model, AIC, BIC
[32]	LSA, TF-IDF, HM, LBP	LBP, MLBP, SURF, HM	Hybrid	REWOA-DBN	DBM, CNN, Autoencoder, RNN, LSTM, DBN
[42]	NLTK, String, re, SentiStrenght, GVAPI	GVAPI, SVM	Hybrid	Negative Binomial Regression w/ log link function	-
[44]	GloVe, Density Matrix, LSTM, NLTK	SIFT, k-means, LSTM	Hybrid	QMN	CNN, FMF, DSEF, MDL, CRNN, h-LSTM, QMSA, DM-CNN, DM-QIMF
[41]	GloVe, Bi-LSTM, MLP	ResNet, MLP	Intermediate	MASA	ATAE-LSTM, IAN, RAM, TNet, MIMN, TomBERT
[43]	GloVe, NLTK, VDCNN	ResNet, DenseNet	Hybrid	AutoML	SVM, RCNN, GBM
[37]	GloVe, MHSA, Bi-GRU	ResNet, Capsule, Attention Mechanism	Attention	EF-Net	ATAE-LSTM, IAN, MemNet, MGAN, Res-MemNet, Res-IAN, Res-MGAN, ESFAN
[48]	FBT+, LDA, LSTM, MLP, MCB+FFT	VGGNet, LSTM, MCB+FFT	Joint/Bilinear	TGANN	LR, SVM, CAN, LightGBM
[34]	BERT	ViT	Attention	CBAN-Add, CBAN-Dot	MVAN, MultiSentiNet, CoMN, FENet, MMHFM, VilBERT, LXMBERT, 2D-Intra-Attention+RoBERTa, VGG + CNN, Relation-Attention, Transformer-Attention
[55]	RoBERTa, MHA	CGM, GCC, VGNN, ResNet	Joint	CoolNet	Res-Target, RAM, AE-LSTM, BERT, MGAN, MIMN, ESAFN, VilBERT, TomBERT, ModalNet-BERT, EF-CapTrBERT, KEF-SaliencyBERT, FITE, HIMT, ITM
[35]	GoogleLens, SentiCircle + ConvNet, GloVe, VADER	GoogleLens, BoVW, LBP, k-means, SVM	Hybrid	HCConvNet-SVM	SVM, Naive Bayesian, KNN, Gradient Boosting
[36]	langdetect, SentEMO	Scraper, SentEMO	-	RoBERTa	-

the fusion process unique. The development of graphs, adopted by Jouyandeh et al. [33] research project, allows each node to be enhanced with associated weight and updated when new data is added, performing Greedy and Random algorithms, to compare elements with the model created for predicting target clients through polarity analysis. The use of attention mechanisms associated with the neural networks created for both fusion and sentiment analysis formulates one of the positive convictions introduced in Qian et al., Cheung et al., Lopes et al. and Gu et al. [34], [37], [43], [48] works as a beneficial practice for increasing the quality of the model's performance and results. However, the way of predicting feelings about the data is one of the conditioning factors in the design of the application proposals and can be individual to each modality [33], [42] (4, 26), during [55] or after [32], [34], [36], [37], [41], [43], [44], [48] the process of concatenating different types of data. Among the models developed are: Text-guided Attention Neural Network (TGANN) [48], with sigmoid and hyperbolic functions for Neural Networks (NN) activation, Adam optimizer and dropout for regulation; Crossmodal Bipolar Attention Network (CBAN) [34], with similarity matrices with softmax function to normalize the weights of each input and cross-entropy loss to balance the model's performance; combination of another matrix structure, known as Dynamic Multi-scale CNN (DM-CNN), for the application of LSTM with sigmoid activation function and dropout strategy [44]; Cross-modal Fine-grained Alignment and Fusion Network (CoolNet) [55] and Deep Belief Network (DBN) [32] for sentiment forecasting, with sigmoid activation function. As an application of comparative methods and to study the performance of the proposals for each contribution, the state-of-the-art models used are: Deep Semantic Network for Multimodal Sentiment Analysis (MultiSentiNet) [34], Multi-Interactive Memory Network (MIMN) and Target-Oriented Multimodal Sentiment Classification (TomBERT) [41], LSTM-Attention [43], Valence Aware Dictionary and Sentiment Reasoner (VADER) [35], [43], among others.

Moreover, the fusion approach in multimodal integration plays a pivotal role in leveraging the complementary information from text and visual modalities. Hybrid approaches, as observed in articles of Gu et al., Mehbodniya et al., Gandhi et al., Zhang et al., Lopes et al., and Kumar et al. [30], [32], [35], [42], [43], [44], combine different techniques to extract features from both modalities before feeding them into the chosen model. These approaches capitalize on the synergistic relationship between textual and visual information, enhancing the overall predictive performance. Intermediate fusion methods, exemplified by Zhou et al. [41], entail processing textual and visual data separately through dedicated models before integrating their representations at an intermediary stage. This approach allows for individualized preprocessing of each modality feature before merging them, potentially capturing nuanced relationships between text and images. On the other hand, attention-based fusion mechanisms, such as those employed in Gu et al. and Cheung et al. [34], [37], dynamically weigh the contributions of textual and visual features during model inference, focusing on the most relevant information for prediction.

This adaptive fusion strategy can effectively handle varying degrees of importance between modalities across different instances, leading to improved predictive accuracy. Other contributions, such as Qian et al. and Xiao et al. [48] and [55], use the joint fusion approach to similarly extract modalities features independently, granting an additional stage to concatenate both extracted features taking into account the performance of mathematical equations specific to the real context of the articles.

In the context of model selection and comparison, various architectures have been proposed to accommodate the multimodal nature of the data and exploit the synergies between text and images. Models like TGANN [48], CBAN [34] and CoolNet [55] integrate advanced neural network components with specialized attention mechanisms to capture intricate relationships within and between modalities. These models leverage sophisticated activation functions, optimization algorithms, and regularization techniques to enhance learning and generalization capabilities. Additionally, ensemble methods like Random Evolutionary Whale Optimization Algorithm - DBN (REWOA-DBN) [32], Quantum-Like Multimodal Network (QMN) [44], and Hybrid Model of CNN with Support Vector Machine (HCConvNet-SVM) [35] aggregate predictions from multiple base models, leveraging the diversity of individual models to improve overall performance. Furthermore, state-of-the-art models such as MultiSentiNet [34] and MIMN [41] serve as benchmarks for evaluating the efficacy of proposed approaches, providing a reference point for assessing performance gains and identifying areas for improvement.

2) TEXT, IMAGE AND AUDIO INTEGRATION ANALYSIS

The development of the authors models in [31], [39], [40], [46], [47], [49], [51], [56], [57], and [60] covers the analysis and fusion of text, audio and video modalities for the study of emotions. The comparative analysis conducted found the different alternatives to deal with visual features, captured via video samples and transformed into singular images [31], [60] or plural sections of visual features [39], [40], [46], [47], [49], [51], [56], [57]. For data extraction and pre-processing, it is possible to find some similarities and differences in the design of these processes and also the tools used. The work of Bi et al. [51] stands out for its different approach to data extraction and pre-processing, documenting three different fusion levels (data layer, feature layer, and decision-making layer), and not specifying which would be used. However, it reveals the use of Automatic Speech Recognition (ASR) for coding audio into text and the contribution of using CNN to extract low and high-level features, similar to [31], [46], and [47]. In addition to this technique, which is also present in Xu et al. [39] and Vepa et al. [57] for text in the first and video for both, there are other techniques common to these and [31], [47], and [56] such as the use of the openSMILE tool to extract audio features. It has capabilities for recognizing and translating clips into text [47], and for generating Low-Level Descriptors (LLDs) and their statistical functions [31], [39], [57]. The articles [40], [56], [57], [58] also mention the use of the Collaborative Voice Analysis Repository (COVAREP) to calculate features for this modality, differentiating it from the [47] which uses the Librosa python library to build the modality vector with Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, Tonal and Centroid values. Also, Xi et al. [49] uses the Spleeter tool with the wav2vec pre-trained model on Chinese common voice datasets. Xu et al. [39], advocating the recognition of emotions individually for each modality vector, relies on SVM to predict the sentiment present in the samples. Shou et al. [46] uses the OpenEar software to get prosody, spectrum and cepstrum variables. The variety of methods applied to visual data, usually done in parallel with the last modality through video analysis, is defined as vast in the methods applied. Facial recognition and general application tasks are developed at the level of frames [46], [47], [57], [58], studying their individuality and dependence on other images taken from videos. Vepa et al., Rahmani et al. and Xu et al. [56], [57], [58] uses the Facet algorithm to identify these features, and in [39], convolutional LSTM generates the embedding sequence with long-term learning of spatial-temporal variables extracted by the Convolutional 3-Dimensional (C3D) model. In Karjee et al. [47], methods such as Mediapipe, OpenCV, and lib are used. Authors from articles [31], [49], [60] all picked a pre-trained model for visual extraction, such as ResNet for [31] and ViT for [49]. Another alternative, implemented by Shou et al. [46], considers the use of Computer Expression Recognition Toolkit (CERT) and Rahmani et al. [56] with Openface2. Lin et al. [40] uses the previous version of OpenFace with Histogram of Oriented Gradients (HOG) algorithm so that both visual and audio features can be stacked by LSTM. However, it is in text extraction that we find the greatest number of prior considerations for specialized contextual preprocessing. Language recognition and translation [39] can be one of the changes to be made to formulate the vocabulary dictionaries [47] or lexical dictionaries [39] needed for Natural Language Processing (NLP). To this end, [47] uses TF-IDF to train the model to extract features, and [39] uses recursion between the CNN and the Recurrent Neural Network (RNN), which receives a vector generated by word2vec to classify the sentiment and update the weights. Also found in [46] and [57] is not only the use of CNN for text embeddings but also Glove for the embeddings of this task. Xi et al. [49] uses Pre-training BERT (PERT), a state-of-the-art Chinese dialogue language model. Lin et al. [40] uses BERT for NLP with self-attention for extraction. Xu et al. [31] uses the pre-trained ELECTRA model to obtain the modality vector. [46] uses the LDA model to capture different topics into prepared collections of words per review, words per topic, and topic per review. [60] uses a pre-trained model with the usage of Convolutional BiGRU.The integration of all the features is carried out through the process of multimodal fusion, some already relying on the sentiment predictions present in each modality [39] and others on representative

TABLE 8. Summary of integrations with textual, visual and audio modalities.

Ref.	Textual Processing	Visual Processing	Audio Processing	Fusion Approach	Models	Baselines
[39]	Google Translator, word2vec, CNN, LSTM	C3D, 3DCNN, ConvLSTM, FC-LSTM	openSMILE, SVM	Hybrid	3DCLS	CNN, RNN, C3D
[40]	BERT, MLP-C	OpenFace, HOG, ERT, CNN, sLSTM, MLP-C	COVAREP, Bi-LSTM, MLP-C	Hybrid	PS-Mixer	LMF, LMFN, ARGF, MFM, RAVEN, MulT, MSAF, MKA, Graph-MFM, Multimodal Graph, GraphCAGE, MFN, MV-LSTM, GATE, AMF-BiGRU, CIA, CIM-MTL, DFF-ATMF
[49]	PERT, Mean Pooling	ViT, Mean Pooling	Spleeter, wav2vec, wav2vec2D	Attention	MTM (MTA+NTA+GP)	MFH, MFB, MLB, Joint-encoding, MulT, EF-MTM w/o res, LF-MTM w/o res, MTM w/o res
[31]	ELECTRA, SnowNLP, word2vec	ResNet	openSMILE	Attention	OpenTranformer	DNN, Audio Transformer, wav2vec, HuBERT, Bi-GRU, Bi-LSTM
[46]	Normalization, Recode, CNN, word2vec, Bi-LSTM	3DCNN	openSMILE, Min-Max Normalization	Attention	DSAGCN	CNN, bc-LSTM, CMN, DialogueRNN, DialogueGCN, AGHMN
[60]	PTWE, Conv-BiGRU, Bi-RNN, Deep CNN	DPTM	librosa	Attention	AMSAER	BLSTM, HMM, DBN, DCGAN, CBP, CMN, ML-SER
[56]	textual LSTM	openFace2, Facet, visual LSTM	COVAREP, openSMILE, acoustic LSTM	Attention	Tree Adaptive Framework	CHF, CAF, MAG, SMM, SEP, CTP, ITP
[57]	Bi-GRU, GloVe, CNN	Co-Attention Matrix, Facets, 3D-CNN	COVAREP, openSMILE	Hybrid	-	MFN, Graph-MFN, DCCA
[51]	Textual LSTM	Visual LSTM	Audio LSTM	Attention	-	CAT-LSTM, Simple LSTM, Hierarchical RNN, Contextual RNN
[47]	TF-IDF	openCV, dib, Mediapipe	openSMILE, pocketsphinx, librosa	Late	Tri-Feature Fusion	Hfusion
[53]	LDA, IBM Watson, GloVe, CNN	CERT	OpenEar	Attention	BLSTM(MA)	MFN, MARN, GME-LSTM(A), TFN, BC-LSTM, EF-LSTM, Bi-LSTM(MA)
[50]	GloVe, BERT	OpenFace2	COVAREP	Hybrid	AOBERT	C-MFN, TFN, LMF, MISA, Graph-MFN, MTMM-ES, TBJE-2, TBJE-3
[58]	GloVe	Facet	COVAREP	Late	CMJRT	EF-LSTM, LF-LSTM, TFN, LMF, MulT, MISA, Self-MM
[54]	-	DenseNet, VGGCNN	SoundNet CNN, FFT	Late	Q-learning algorithm	SWM, RF, LSTM

vectors [47], [51], [57]. Except for Xu et al. [39], which defines the fusion process by using the MLK algorithm and Radial Basis Functions (RBF) kernel to fuse the emotional vector of each modality, all the others highlight the use of the attention mechanism before fusion [40], [57], [58], during fusion [47], [51], [58] or both approaches combined [31], [49], [60], either employing a fusion kernel [57], by neural networks defined for sentiment analysis [40], [47] or joint representation [49], [58]. They explore the advantages of this approach, with [51], [57] exploiting self and cross attention to learn long-term dependencies with CAT-LSTM as a comparative competition to the performance of the Hierarchical RNN and Simple LSTM models. Xi et al. and Xu et al. [49], [58] considers a joint representation approach, modulating modalities into personalized multimodal pairs. For comparable NN configuration analysis, with the ReLU [31], [40], [46], [47], [49], [57], [60], softmax [47], [58] and sigmoid [56] activation functions, as well as optimizers such as Adam [31], [46], [49], [57], [60] and Stochastic Gradient Descent (SGD) [39], [40].

Presented in Chen et al. work [54], both audio and visual modalities provide the foundation for the development of a system that uses emotion recognition through a multimodal approach for price analysis on an e-commerce platform. Images and audio samples are obtained by granular decomposition of videos of customer behavior and opinions. For audio, the SoundNet CNN application, noted in [54] for its good learning ability, condenses knowledge on MFCC and other features at the level of the frames identified, followed by an algorithmic transformation using the Fast Fourier Transformer (FFT). The images, on the other hand, are constructed in the form of RGB-based frames and also extracted using two types of neural networks: Dense Connection CNN (DenseNet CNN) and Visual Geometry Group (VGG) to recognize sentiment singular (image) and plural (scene) granularity of the variable [54]. The selection of classifiers relies on the application of a specific RNN, LSTM with batch normalization layer to submit the fusion of the models of each modality according to weight voting.

The fusion approach encompasses various strategies utilized to integrate textual, visual, and audio modalities for emotion analysis. Hybrid fusion emerges as a prevalent approach, evident in models such as Polar and Strength Vector Mixer Model (PS-Mixer) [40] and OpenTranformer [31]. These models integrate features from different modalities before feeding them into the model, thus enabling beneficial supplementary information provided by each modality. Additionally, attention mechanisms play a significant role in dynamically weighting the contributions of each modality during model inference. Models such as DSAGCN [46] and Multi Task Model (MTM) [49] utilize attention mechanisms to selectively prioritize relevant features from disparate modalities, thereby facilitating adaptive integration following task requirements. In contrast, late fusion approaches, as exemplified by models such as Cross-Modal Joint Representation Transformer (CMJRT) [58], incorporate predictions derived from individual modalities at a later stage of the model's architectural development. This strategy allows the model to process information from each modality independently before combining them to make final predictions, potentially capturing complex interactions between modalities. The models display a diverse array of neural network architectures and algorithms employed to process data from textual, visual, and audio modalities. Advanced architectures such as CNNs [39], [46], LSTMs [31], [46], [47], [50], [51], [54], [58], [60], Bi-GRUs [31], and transformer variants [31], [40], [46], [49], [50], [58], [60] are leveraged to capture intricate relationships within and between modalities. For instance, models such as OpenTranformer [31] employ Transformer architectures with attention mechanisms to process multimodal inputs and make predictions. On the other hand, models such as DSAGCN [46] employ CNNs and LSTMs regarding the potential of these resources for both visual and auditive feature extraction, respectively, but also for the formal comparison of the employed model as common baselines. Other traditional architectures such as RNN and C3D serve as reference models against which the proposed approaches are evaluated. These baseline models provide a comparative reference point for assessing the effectiveness of proposed approaches in multimodal emotion analysis tasks. By comparing against established benchmarks, researchers can gauge the performance improvements achieved by fusion strategies and advanced model architectures in handling multimodal data.

3) E-T, EEG AND VISUAL INTEGRATION ANALYSIS

The computational integration of biometric data and its analysis through AI has revolutionized research on human perceptions and how to generate knowledge that is more in line with reality. In agreement with [45] and [59], the multimodal approach to signals obtained by capturing eye movements and EEG, stimulated by images of different product designs, makes it possible to quantify and study consumer preferences. Its use is restricted by the initial stage of capturing information, which requires the acquisition of devices embedded with software specialized in transforming these activities into signals [45] - Curry 8, Tobii Pro Lab, among others. Reference [59] considers these variables through EEG and EOG signals which, processed separately, are represented by characteristics associated with their signal and distribution metrics. These are obtained respectively using techniques such as the Short Time Fourier Transformer (STFT) and Electrooculograph power density estimation (EOG-PDE) to make the signal more robust for pattern recognition. In [45], EEG signals and E-T characteristics (e.g. fixation, saccade) are submitted to pre-trained continuous emotion recognition models and fuzzy inference systems. They also highlight their interest in evaluating valence and arousal as independent variables of consumer preference, assessed by the attention mechanism that consolidates temporal features of the Spatial-Temporal Neural Network model, which the authors have named Att-2D-CNN. Its use of fuzzy systems guarantees the perception of general preferences and evaluates with the proposed model whether the orientation is common with the polarity of the predicted feeling. The non-stationary and non-linear nature of both signals requires the need for data normalization in both contributions, except personalization in [59] to keep the original characteristics of the signal as present as possible. Techniques applied to neural networks such as backpropagation and the SGD method are also used to update the weights and increase the flexibility and scalability of the proposed models.

The methods used to combine data from visual, EEG and Eye-Tracking (E-T) modalities for different applications are displayed in the Fusion Approach column. One common technique is hybrid fusion, which combines characteristics from several modalities at various points in the model design. For example, Wu et al. [59] used a hybrid fusion strategy to examine EEG data in addition to visual inputs by combining pattern recognition algorithms with EEG signal processing techniques. Comparably, Zhu et al. [45] use fuzzy algorithms in a hybrid fusion technique to combine characteristics taken from eye-tracking data with EEG signals. By using this method, the model may better capture intricate patterns and connections by utilizing complementary information from several modalities.

To interpret data from EEG, E-T, and visual modalities, a range of neural network designs and signal processing approaches are shown by the selected models. A Deep Graph Neural Network (DGNN) is used in Wu et al. [59] to assess EEG data and visual patterns at the same time. This model architecture is intended to efficiently process visual information and capture temporal and spatial relationships included in EEG data. Similar to this, Zhu et al. [45] evaluate data from EEG and eye-tracking modalities using an Attention-based 2D Convolutional Neural Network (Att-2D-CNN). These models perform better in multimodal analysis tasks by focusing on pertinent characteristics and patterns within the input data by integrating attention processes. Conventional machine learning techniques and well-known neural network designs are used as reference models in the baselines to assess the suggested methods. For instance, baseline models such as Artificial Neural Networks (ANNs), SqueezeNet, GoogleNet, ResNet, DarkNet, Inception, and Inception-ResNet are utilized to compare the performance of the suggested DGNN architecture in the study by Wu et al. [59]. These baseline models serve as a standard by which to evaluate how well the suggested fusion techniques and model architectures analyze multimodal data. Similar to this, Zhu et al. [45] highlight the advantages of their suggested method by contrasting their Att-2D-CNN model with SVM, Random Forests (RFs), Principal Component Regression Neural Networks (PCRNNs),

TABLE 9. Summary of integrations with visual, EEG and E-T modalities.

Ref.	EEG Processing	E-T Processing	Visual Processing	Fusion Approach	Models	Baselines
[59]	Normalization, SSIM, E, C, HOC, DFT, PSD, STFT, EEG-Lab	EOG-PDE, PSD, CGF, F.U., RMSF, PSE, IDF, EYE-EEG	Pattern Recognition	Hybrid	DGNN	ANN, SqueezeNet, GoogleNet, ResNet, DarkNet, Inception, Inception-ResNet
[45]	Att-2D-CNN, FIS, Curvy 8, Neuroscan electrode cap, SynAmps RTamplifier	Tobii Pro Glasses 2	Fuzzy Systems	Hybrid	Att-2D-CNN	SVM, RF, PCRNN, Att-Bi-LSTM

and Attention-based Bi-Directional Long Short-Term Memory Networks (Att-Bi-LSTMs).

E. ETHICAL AND REGULATION CONSIDERATIONS

Of the 35 articles returned, Neuwirth et al and Diaz-Rodriguez et al. [25], [26] correctly list the ethical and regulatory considerations in the application of AI technologies, taking into account government proposals such as the AI Act and the principles of creating trustworthy AI. This also illustrates the lack of ethical and legal concerns on the remaining papers, unveiling the crucial need to document the procedures required for an accurate AI system development [25]. Even though Marketing is one of the central investigation areas, mentioned in various sections in [25] and [26], the purpose of both authors aims to present the appropriate deliberations across the multiple universes where the application of AI is found. Confirming that these criteria apply to the design, development, and implementation of these systems to the public will help AI to continue along paths that are beneficial to the evolution of humanity [25], [26].

AI is currently seen as a great and valuable way of overcoming any previously immeasurable and unattainable task, challenge, or objective in any business area, making it a powerful and indispensable tool [26]. However, it is also subject to various studies and investigations to ascertain whether the benefits of AI can outweigh the various risks to which they are exposed in their adoption in the short, medium, and long term. Although both authors [25], [26] find flaws in the legislation at a global level, there are disparate governmental indications in each world power, also differing from the perspective of free access to AI systems that may not coincide with the restrictions in force in a given geographical area. The most scrutinized of the two, the AI Act, was presented for the first time in April 2021 by the European Union, with a group of articles regulating the use of the intelligence generated by these innovations [25]. Both authors [25], [26] guarantee, throughout their structure, the study of possible formulations of structures that violate human rights such as the privacy and security of the actors themselves and third parties, reviewed in more detail below. The authors highlight the fact that the European Union's proposal, even though it only covers its geographical area, represents a major step forward in the construction of regulations [25] that will make it possible to control and direct AI worldwide in areas where the risk can be measured and mitigated without developing new types of adversity [25], [26]. In addition to the AI Act, other frameworks for legal and ethical action on associated resources are also referenced to make it a standard requirement for all [26]. Both teams [25], [26] also highlight the repercussions caused by insufficient answers to these questions, which negatively evaluate the systems. Among the reasons is the difficulty in recognizing the integrity of all the characteristics that build the model and make its outcome relevant [26]. Attention to these points of variability in the response to ethical and regulatory considerations defines these articles as crucial contributions to understanding the limitations that ensure trust in AI applications [25], [26].

In Neuwirth et al. document [25], it is possible to find a description of the main objectives of the AI Act and how it is possible to classify all procedures that include the presence of intelligent systems. This contextualization leads to a detailed explanation of the different high-risk procedures accompanied by examples of current practices that are considered unacceptable in terms of their use. The individual factors of each type of system are highlighted by the various contours in their structure that challenge a worrying number of ethical and legal considerations [25]. These are aggravated when it is possible to assess their impact not only individually but also jointly, triggered by the successive continuation of dangerous practices unknown to their users [25], [26]. Neuwirth et al. [25] translate each scenario into a clear example of existing systems that reproduce these same practices, highlighting the lack of monitoring and demanding that it be carried out correctly. In all of these interactions, vulnerabilities in compliance with ethical considerations are uncovered, reflecting the need to standardize the process. Still on this last objective, the authors conclude that the delay in achieving it guarantees irreversible consequences if its priority is not equated with other situations of greater tension and concern such as the emergence of conflicts, wars, and similar crises [25].

Diaz-Rodriguez et al. [26], on the other hand, provide more exploratory and comprehensive documentation of all the characteristics that need to be analyzed and confirmed so that the reliability of an AI system can be fully guaranteed. This same certification can be awarded to systems that have already been made available to the public or are yet to be made available. To this end, the authors [26] emphasize the elements and how the dependence between the pillars and the requirements makes it possible to award a reliability label to an artificial intelligence system. Following the guidelines presented in the AI Act proposal but also recognizing other recommendations developed by researchers and other corporations, they developed a study specifying the importance of each topic for maximizing this classification [25], [26]. Through the first categorizations made available by the AI Act, they provide a critical analysis of the principles that drive the development of artificial intelligence, a philosophical approach to the ethical considerations associated with AI, how current regulations approach AI with an associated risk approach and what criteria should be in place for the system under construction. The statement of all the necessary pieces for building a reliable system then allows the monitoring cycle to be defined to ensure that all practices are maintained, even if they are high-risk, followed by a set of guidelines that keep them in line with the law [25], [26].

Although the approaches to the ethical and regulatory considerations of AI taken by the two research teams [25], [26] are different, they complement each other in revealing the circumstances and minimum requirements for use and maintenance. In this way, contextualizing the issue of multimodality in sentiment analysis is a challenging problem with many nuances to consider. As described in both articles, but more emphasized by Neuwirth et al. [25], the capture of biometric data in real-time is one of the unacceptable procedures because it puts the customer's privacy at risk due to its possible identification for the self-interest of the interested parties. Although this type of situation is difficult to identify in large-scale models [26], it requires a constant description of the elements and variables used to form the knowledge for the model. In this way, multimodality is required to be anonymous about all the data that represents the interactions of the different customers, fulfilling privacy, non-discrimination, and mitigating the possible creation of customer credit systems [25], [26]. The care taken in choosing the heterogeneity of the data lies exclusively with those responsible for its adoption and modeling, without any of them allowing the system to acquire more information than it is intended for. Diaz-Rodriguez et al. [26] also mention that all types of studies on the impact of each modality should be reported to streamline different studies in terms of their granularity and show the system's performance.

In addition to the above-mentioned considerations, it is also necessary to constantly update the data to prevent the results from being impacted by biases or other types of injustice due to the redundancy of the model's performance. In the case of sentiment analysis for an environment where Marketing campaigns are expressed, the issue of bias warrants other concerns regarding the use of this information as a way of predicting the following. According to Neuwirth et al. [25], concerning emotion recognition, it is crucial to find an adjustment time interval so that the current state of customers will not be the same in the future, triggered by the dynamic and unpredictable interaction of customer behaviors. In agreement with Diaz-Rodriguez et al. [26], both relate that any type of behavior manipulated or used to identify the customer should be suppressed to guarantee the robustness and security of all those involved.

Given the pillars and requirements presented that allow an AI system to be reliable with its activity provided to the general public [26], the ethical and regulatory considerations for Marketing encounter several challenges to be overcome [25]. The process must include a rigorous study of the associated risk throughout the system's life cycle, ensuring that all details are reported. It must also make use of quality-assured and certified data resources so that the analysis of the activity produced is perceptible. Documentation of these items must be provided regularly so that the understanding of all developments is monitored and supervised. This, in turn, must always be carried out while ensuring that logical and human-centered considerations are maintained. The application of multimodality on sentiment analysis to today's consumers is achievable by maintaining the good practice of all the measures advocated by current legislation.

F. CHALLENGES AND OPPORTUNITIES FOR POTENTIAL FUTURE APPLICATIONS

Exploring different applications of Multimodal Sentiment Analysis for Marketing imperatively needs to recognize a series of existing challenges and possible promising opportunities for the future of this approach [25], [26], [52], [61]. As it is a new way of working, it acquires a diversity of interpretations considering the real case in question [33], [52]. However, the uniqueness of each situation unveils new restrictions and the need to find hypotheses capable of preserving the advantages that AI currently provides.

The limitations analogous to the programs previously reported are part of different development phases, maximizing the possibility of numerous points of failure or inaccuracy [25], [26], [61]. The design of a reliable system relies on several decisions on the selection, representation, and fusion methods of data that are significant to the problem in question, creating external factors that influence non-controllable changes [25], [52]. In this way, data integration significantly impacts the following systems' performance, as one of the major problems in fulfilling a proper shape [34], [43], [49], [56]. Not knowing the nature of crucial processes such as the collection of data, the analysis and transformation of the input samples made available to the models compromises the guarantee of their effectiveness due to the lack of quality [42], [45], [48], [61]. This can be caused by the granularity of each modality, the individual and multiple organization of the set, or even the annotations connected to each unit, compromising the benefits of fusing modalities [32], [38], [48]. Thus, this type of obstacle can lead to arguments that compromise the overall investigation, encountering inefficient multimodal adoptions when compared to the traditional ones of just one modality [30]. Documenting all the details that redefine the model structure is crucial to ensure all the impacts that could affect the obtained outcomes, ensuring coherency with the ethical and regulatory recommendations mentioned above [25], [26], [61].

In addition to the obstacles mentioned, which may be part of the subsequent consequences, it is also important to note the lack of transparency in the perception of the work carried out by the model [43], [49], [57], [59], [61]. The complexity inherent in understanding its procedures jeopardizes a clear interpretation capable of comparing two similar cases and understanding their disparities, even if they are superfluous [25], [26], [43], [61]. As well as the involvement of AI algorithms and methods adding to the computational and time costs, it also raises vulnerabilities in terms of the confidence given to the results obtained and the knowledge generated [33], [43], [49], [59]. This issue is compounded by the aggravation of using emotional expressions, which raises additional concerns about their correct generalization and contextualization. As the sharing of this type of information is inconsistent and difficult to moderate, its scalability compromises its application in real-time, without ethical and structural considerations being ruled out [33], [47], [51], [61]. As with the limitations mentioned above, the similarities in their application and use depend not only on the ethics of all those involved but also on the regulations to be followed [25], [26]. Although data privacy is a general and obligatory condition to be maintained, there are nuances to the whole conception and application of this type of system that compromise a future outlook equal to the current state [34], [61]. The need to legalize certain procedures and restrict others places the application of AI as an unwise practice without first formulating all these considerations [52], [61].

To overcome these obstacles, the spotlight of today is focused on the incentive to formulate new promises for the future use of Affective Computing and Multimodal AI [42], [43], [61]. The common perspective of those who adopt it translates into a panoply of tasks that were previously unthinkable or difficult for the human hand to achieve, such as the constant monitoring of the brand's image and reputation and its presence on the market [30], [42], [61]. The care required to control the form and essence of established advertising also makes it possible to manage and predict current consumers, using their experiences to attract new customers [33], [48], [51]. Knowing current trends leads to a flexible application to the requirements of each sector of activity [33], [47], estimating the coefficients that are beneficial to the associated competitiveness. In this way, multimodality allows the creation of new content to be imbued with stimuli enhanced by feelings or polarities that attract new consumers and retain those already acquired [35], [43]. The advantage gained by managing the dependencies of this analytical process promotes adaptations in the interfaces made available, ideally designed without the presence of discrimination or vulnerabilities [25], [26]. The combination of other technologies, such as Virtual Reality or Augmented Reality, offers new paths where multimodality is contained within easily mapped and personalized parameters [47].

IV. CONCLUSION

In conclusion, this research undertakes the exploration of the integration of a multimodal perspective into sentiment analysis for enhancing marketing strategies, conducted as part of the PHYNHANCAI project. Aiming to provide customized solutions for the phygital world of marketing strategies, the incorporation of sentiment analysis is proposed to gain insights from various measurable modalities, contributing to a better understanding of current public trends. The project sets out to bridge the gap between traditional unimodal sentiment analysis and the emerging field of multimodal sentiment analysis, focusing on various modalities, including text, images, audio, and physiological signals.

The methodology chosen for the research involves the application of the PRISMA methodology for the systematic review of scientific contributions. As the project progresses, it is anticipated that the findings will contribute to the development of optimized solutions for sentiment analysis in marketing. The ethical and legal considerations associated with the use of multimodal AI in marketing will also be explored to ensure responsible and transparent practices.

Pursuing the goal of harnessing the importance of multimodal affective computing, not only for marketing applications but for other research areas, it is crucial to examine the constraints encountered during this systematic research. Due to that, this chapter documents provides insights into the main areas that warrant concerns, presenting alternatives for further improvement.

A. METHODOLOGICAL SEARCH

To perform a systematic review, rigorous methodological research items are essential to leverage the potential of affective computing for marketing. This involved a comprehensive exploration of existing methodologies and approaches in the literature. An analysis of various techniques and frameworks used in other systematic reviews identifies the PRISMA methodology as the best suited for this topic due to the robustness of reproducing the whole process. Even creating a solid foundation to find valuable contributions, there are still some metadata issues that may change the display of articles and alter conclusions direction. Similarly to this, as mentioned in [4], new and contextual vocabulary related to the research theme can sometimes limit the acquisition of relevant articles to the research. Special concepts with recent origin by uniting two or more words urge the need to keep up with the latest trends and ontology, giving credit to this kind of work for its contribution.

B. EXPERIMENTAL ENVIRONMENTS FOR AFFECTIVE COMPUTING

The experimental environments employed for the study of affective computing play a pivotal role in shaping the outcomes of this complex field. Even detailing the settings in which affective computing analysis was conducted, there is a great number of aspects that could dismiss the effectiveness of observed findings. The first one is related to the laboratory stages preceding the capture of data, giving a simulated input of affective cues to build a trustworthy dataset. Still connected to this, the second one regards adapting real-world scenarios with particular implications to study affective responses. The ontology used for affective computing mixes the definition of emotion recognition, sentiment analysis, and opinion mining, imposing undercover boundaries to the obtained results. Understanding the experimental environments is crucial for interpreting the practical implications of research within the dynamic landscape of marketing.

C. ETHICAL AND REGULATIVE RESPONSIBILITY

Our commitment to ethical and regulative responsibility is fundamental to the integrity of our research. This section addresses the ethical considerations that guided our decision-making throughout the study. We explore the principles and guidelines followed to ensure the responsible use of Affective Computing technologies. Additionally, we emphasize the geographical dimension of our contributions, acknowledging the importance of considering diverse locations in the ethical framework. By doing so, we contribute to a nuanced understanding of the global impact and responsible deployment of affective computing in marketing.

REFERENCES

[1] C. Grange, I. Benbasat, and A. Burton-Jones, "A network-based conceptualization of social commerce and social commerce value," *Comput. Hum. Behav.*, vol. 108, Sep. 2020, Art. no. 105855.

[2] C. Constantinescu, B. Pokorni, and J. Wimmer, "Affective production systems: Foundations, reference model and roadmap for implementation and validation," *Proc. CIRP*, vol. 104, pp. 1783–1786, 2021.

[3] L. McShane, E. Pancer, M. P. Deng, and Q. Emoji, "Playfulness, and brand engagement on Twitter," *J. Interact. Marketing*, vol. 53, pp. 96–110, Aug. 2021.

[4] P. Del Vecchio, G. Secundo, and A. Garzoni, "Phygital technologies and environments for breakthrough innovation in customers' and citizens' journey. A critical literature review and future agenda," *Technological Forecasting Social Change*, vol. 189, Apr. 2023, Art. no. 122342.

[5] Q. Deng, Y. Wang, M. Rod, and S. Ji, "Speak to head and heart: The effects of linguistic features on B2B brand engagement on social media," *Ind. Marketing Manage.*, vol. 99, pp. 1–15, Nov. 2021.

[6] J. Serrano-Guerrero, F. P. Romero, and J. A. Olivas, "Fuzzy logic applied to opinion mining: A review," *Knowl.-Based Syst.*, vol. 222, Jun. 2021, Art. no. 107018.

[7] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jul. 2020.

[8] A. Borawska and M. Łatuszyńska, "The use of neurophysiological measures in studying social advertising effectiveness," *Procedia Comput. Sci.*, vol. 176, pp. 2487–2496, 2020.

[9] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134.

[10] R. Chaturvedi, S. Verma, R. Das, and Y. K. Dwivedi, "Social companionship with artificial intelligence: Recent trends and future avenues," *Technological Forecasting Social Change*, vol. 193, Aug. 2023, Art. no. 122634.

[11] D. W. Prabowo, H. A. Nugroho, N. A. Setiawan, and J. Debayle, "A systematic literature review of emotion recognition using EEG signals," *Cognit. Syst. Res.*, vol. 82, Dec. 2023, Art. no. 101152.

[12] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, Mar. 2023.

[13] W. Li, Z. Zhang, and A. Song, "Physiological-signal-based emotion recognition: An Odyssey from methodology to philosophy," *Measurement*, vol. 172, Feb. 2021, Art. no. 108747.

[14] I. César, I. Pereira, F. Rodrigues, V. Miguéis, S. Nicola, and A. Madureira, "Multimodal learning applications on digital marketing: A review," in *Proc. 23rd Int. Conf. Hybrid Intell. Syst. (HIS)*, Porto, Portugal, 2023, pp. 1–10.

[15] P. Singh Tomar, K. Mathur, and U. Suman, "Unimodal approaches for emotion recognition: A systematic review," *Cognit. Syst. Res.*, vol. 77, pp. 94–109, Jan. 2023.

[16] S. B. Daily, M. T. James, D. Cherry, J. J. Porter, S. S. Darnell, J. Isaac, and T. Roy, "Chapter 9—Affective computing: Historical foundations, current applications, and future trends," in *Emotions and Affect in Human Factors And Human-Computer Interaction*. San Diego, CA, USA: Academic, 2017, pp. 213–231, doi: 10.1016/B978-0-12-801851-4.00009-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128018514000094

[17] R. Arya, J. Singh, and A. Kumar, "A survey of multidisciplinary domains contributing to affective computing," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100399.

[18] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *J. Netw. Comput. Appl.*, vol. 149, Jan. 2020, Art. no. 102447.

[19] S. C. Leong, Y. M. Tang, C. H. Lai, and C. K. M. Lee, "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing," *Comput. Sci. Rev.*, vol. 48, May 2023, Art. no. 100545.

[20] R. K. Behera, P. K. Bala, N. P. Rana, and H. Kizgin, "Cognitive computing based ethical principles for improving organisational reputation: A B2B digital marketing perspective," *J. Bus. Res.*, vol. 141, pp. 685–701, Mar. 2022.

[21] L. Ma and B. Sun, "Machine learning and AI in marketing–connecting computing power to human insights," *Int. J. Res. Marketing*, vol. 37, no. 3, pp. 481–504, Sep. 2020.

[22] S. Polevikov, "Advancing AI in healthcare: A comprehensive review of best practices," *Clinica Chim. Acta*, vol. 548, Aug. 2023, Art. no. 117519.

[23] I. Mezgár and J. Váncza, "From ethics to standards—A path via responsible AI to cyber-physical production systems," *Annu. Rev. Control*, vol. 53, pp. 391–404, Jun. 2022.

[24] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "Decoding ChatGPT: A taxonomy of existing research, current challenges, and possible future directions," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 35, no. 8, Sep. 2023, Art. no. 101675.

[25] R. J. Neuwirth, "Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA)," *Comput. Law Secur. Rev.*, vol. 48, Apr. 2023, Art. no. 105798.

[26] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101896.

[27] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[28] C. Sohrabi, T. Franchi, G. Mathew, A. Kerwan, M. Nicola, M. Griffin, M. Agha, and R. Agha, "PRISMA 2020 statement: What's new and the importance of reporting guidelines," *Int. J. Surgery*, vol. 88, Apr. 2021, Art. no. 105918.

[29] M.-H. Huang and R. T. Rust, "A framework for collaborative artificial intelligence in marketing," *J. Retailing*, vol. 98, no. 2, pp. 209–223, Jun. 2022.

[30] W. Gu, K. W. Chan, J. Kwon, C. Dhaoui, and F. Septianto, "Informational vs. emotional B2B firm-generated-content on social media engagement: Computerized visual and textual content analysis," *Ind. Marketing Manage.*, vol. 112, pp. 98–112, Jul. 2023.

[31] W. Xu, X. Zhang, R. Chen, and Z. Yang, "How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce," *Decis. Support Syst.*, vol. 172, Sep. 2023, Art. no. 113984.

[32] A. Mehbodniya, M. V. Rao, L. G. David, K. G. Joe Nigel, and P. Vennam, "Online product sentiment analysis using random evolutionary whale optimization algorithm and deep belief network," *Pattern Recognit. Lett.*, vol. 159, pp. 1–8, Jul. 2022.

[33] F. Jouyandeh and P. M. Zadeh, "IPARS: An image-based personalized advertisement recommendation system on social networks," *Proc. Comput. Sci.*, vol. 201, pp. 375–382, 2022.

[34] T.-H. Cheung and K.-M. Lam, "Crossmodal bipolar attention for multimodal classification on social media," *Neurocomputing*, vol. 514, pp. 1–12, Dec. 2022.

[35] A. Kumar, K. Srinivasan, W.-H. Cheng, and A. Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data," *Inf. Process. Manage.*, vol. 57, no. 1, Jan. 2020, Art. no. 102141.

[36] L. De Bruyne, A. Karimi, O. De Clercq, A. Prati, and V. Hoste, "Aspect-based emotion analysis and multimodal coreference: A case study of customer comments on adidas Instagram posts," in *Proc. 13th Lang. Resour. Eval. Conf.*, Jun. 2022, pp. 574–580.

[37] D. Gu, J. Wang, S. Cai, C. Yang, Z. Song, H. Zhao, L. Xiao, and H. Wang, "Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021.

[38] H. Li, H. Ji, H. Liu, D. Cai, and H. Gao, "Is a picture worth a thousand words? Understanding the role of review photo sentiment and text-photo sentiment disparity using deep learning algorithms," *Tourism Manage.*, vol. 92, Oct. 2022, Art. no. 104559.

[39] G. Xu, W. Li, and J. Liu, "A social emotion classification approach using multi-model fusion," *Future Gener. Comput. Syst.*, vol. 102, pp. 347–356, Jan. 2020.

[40] H. Lin, P. Zhang, J. Ling, Z. Yang, L. K. Lee, and W. Liu, "PS-mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103229.

[41] J. Zhou, J. Zhao, J. X. Huang, Q. V. Hu, and L. He, "MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis," *Neurocomputing*, vol. 455, pp. 47–58, Sep. 2021.

[42] M. Gandhi and A. K. Kar, "How do fortune firms build a social presence on social media platforms? Insights from multi-modal analytics," *Technological Forecasting Social Change*, vol. 182, Sep. 2022, Art. no. 121829.

[43] V. Lopes, A. Gaspar, L. A. Alexandre, and J. Cordeiro, "An AutoML-based approach to multimodal image sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–9.

[44] Y. Zhang, D. Song, X. Li, P. Zhang, P. Wang, L. Rong, G. Yu, and B. Wang, "A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Inf. Fusion*, vol. 62, pp. 14–31, Oct. 2020.

[45] S. Zhu, J. Qi, J. Hu, and S. Hao, "A new approach for product evaluation based on integration of EEG and eye-tracking," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101601.

[46] Y. Shou, T. Meng, W. Ai, S. Yang, and K. Li, "Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis," *Neurocomputing*, vol. 501, pp. 629–639, Aug. 2022.

[47] J. Karjee, G. Dwivedi, A. Bhagavath, and P. Ranjan, "A lightweight multimodal learning model to recognize user sentiment in mobile devices," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2023, pp. 1–6.

[48] Y. Qian, W. Xu, X. Liu, H. Ling, Y. Jiang, Y. Chai, and Y. Liu, "Popularity prediction for marketer-generated content: A text-guided attention neural network for multi-modal feature fusion," *Inf. Process. Manage.*, vol. 59, no. 4, Jul. 2022, Art. no. 102984.

[49] D. Xi, L. Tang, R. Chen, and W. Xu, "A multimodal time-series method for gifting prediction in live streaming platforms," *Inf. Process. Manage.*, vol. 60, no. 3, May 2023, Art. no. 103254.

[50] K. Kim and S. Park, "AOBERT: All-modalities-in-one BERT for multimodal sentiment analysis," *Inf. Fusion*, vol. 92, pp. 37–45, Apr. 2023.

[51] W. Bi, Y. Xie, Z. Dong, and H. Li, "Enterprise strategic management from the perspective of business ecosystem construction based on multimodal emotion recognition," *Frontiers Psychol.*, vol. 13, pp. 1–11, Mar. 2022.

[52] D. Grewal, D. Herhausen, S. Ludwig, and F. V. Ordenes, "The future of digital communication research: Considering dynamics and multimodality," *J. Retailing*, vol. 98, no. 2, pp. 224–240, Jun. 2022.

[53] Z. Wang, P. Gao, and X. Chu, "Sentiment analysis from customer-generated online videos on product review using topic modeling and multi-attention BLSTM," *Adv. Eng. Informat.*, vol. 52, Apr. 2022, Art. no. 101588.

[54] J. Chen, Z. Zhong, Q. Feng, and L. Liu, "The multimodal emotion information analysis of e-commerce online pricing in electronic word of mouth," *J. Global Inf. Manage.*, vol. 30, no. 11, pp. 1–17, Dec. 2022.

[55] L. Xiao, X. Wu, S. Yang, J. Xu, J. Zhou, and L. He, "Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis," *Inf. Process. Manage.*, vol. 60, no. 6, Nov. 2023, Art. no. 103508.

[56] S. Rahmani, S. Hosseini, R. Zall, M. R. Kangavari, S. Kamran, and W. Hua, "Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects," *Knowledge-Based Syst.*, vol. 261, Feb. 2023, Art. no. 110219.

[57] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4477–4481.

[58] M. Xu, F. Liang, X. Su, and C. Fang, "CMJRT: Cross-modal joint representation transformer for multimodal sentiment analysis," *IEEE Access*, vol. 10, pp. 131671–131679, 2022.

[59] Q. Wu, N. Dey, F. Shi, R. G. Crespo, and R. S. Sherratt, "Emotion classification on eye-tracking and electroencephalograph fused signals employing deep gradient neural networks," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107752.

[60] A. Aslam, A. B. Sargano, and Z. Habib, "Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks," *Appl. Soft Comput.*, vol. 144, Sep. 2023, Art. no. 110494.

[61] A. Capatina, M. Kachour, J. Lichy, A. Micu, A.-E. Micu, and F. Codignola, "Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations," *Technological Forecasting Social Change*, vol. 151, Feb. 2020, Art. no. 119794.

INÊS CÉSAR was born in Porto, Portugal, in 2000. She received the B.S. degree in computer sciences from the Polytechnic of Porto, Porto, in 2022, where she is currently pursuing the M.S. degree in information and knowledge systems. She has been participating in the PHYNHANCAI Project as a Research Student with the Interdisciplinary Studies Research Center, Polytechnic of Porto, since 2023. She has already published three scientific articles about her main scientific areas, related to data science, computer science, affective computing, and customer relationship management.

IVO PEREIRA was born in Porto, Portugal, in 1984. He received the B.Sc. and M.Sc. degrees in informatics engineering from the Institute of Engineering—Polytechnic of Porto, in 2006 and 2009, respectively, and the Ph.D. degree in electronics and computers engineering from the University of Tras-os-Montes and Alto Douro, in 2014. He is currently an Assistant Professor with the University Fernando Pessoa, Porto, a Senior Researcher with the Interdisciplinary Studies Research Center (ISRC), Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), and a Senior Researcher with E-goi. He has participated in more than seven research and development projects. He is currently a principal investigator of the PHYNHANCAI Project. He has published more than 65 scientific papers in international conference proceedings, journals, and book chapters.

FÁTIMA RODRIGUES received the B.Sc. degree from the University of Minho, in 1989, the M.Sc. degree from the University of Porto, Portugal, in 1997, and the Ph.D. degree in computer science from the University of Minho, in 2000. She has been a Professor with the Department of Computer Science, Institute of Engineering, Polytechnic of Porto, since 1993. Her main scientific and research interests include data science, data warehouses, recommendation systems, process automation, and optimization for decision systems.

VERA L. MIGUÉIS received the Ph.D. degree in industrial engineering and management from INESC TEC. She is currently an Associate Professor with the Department of Industrial Engineering and Management, Faculty of Engineering, University of Porto (FEUP), Portugal. She is also the Director of the Undergraduate Program in Industrial Engineering and Management, FEUP, and integrates the Executive Committee of the Department. She is also a Researcher with INESC TEC. Her research interests include data analysis and quantitative methods to support the decision-making process. She has mainly worked on educational data mining and analytical customer relationship management. She has published articles in several international computer science journals.

SUSANA NICOLA received the bachelor's degree in mathematics, the M.Sc. degree in sciences and engineering from the University of Porto, and the Ph.D. degree in engineering and industrial management. She has been a Professor with the Mathematics Department, ISEP, since 1997. As an Integrated Member of INESCTEC. Her research interests include innovation, entrepreneurship, value creation, low-code technology, innovative pedagogical methods, and fuzzy theory. She has received awards in pedagogical innovation, innovation, entrepreneurship, and develops disruptive disciplines with ISEP, particularly in process digitalization and entrepreneurship, involving companies as mentors. She is currently a Co-PI the PHYNHANCAI Project and coordinated the DRIVE-MATH. She is also a member of the Project Management and Coordination Team of the e-H4F Project. She is also the Director of the Digital BIP Sustainable Manufacturing Program, ISEP. She has supervised Ph.D., M.Sc. students, and internship programs. She is recognized as a Professional Engineering Educator by ENTER.

ANA MADUREIRA (Senior Member, IEEE) received the B.S. degree in computer science engineering, the M.S. degree in electrical and computers engineering, the Ph.D. degree in production and systems, the Habilitation degree in informatics, in 1993, 1996, 2003, and 2022, respectively. She is currently a Coordinator Professor with Habilitation with ISEP/P.PORTO and has a significant experience in teaching computer science undergraduate and graduate courses and in master's and Ph.D. thesis supervision, since 1994. She is also the Director of Research and Development Interdisciplinary Studies Research Center (ISRC). She has been involved in more than 20 research and development projects from which she was a principal investigator of four research and development projects, mainly on the AI based developing of intelligent scheduling systems. She has a patent published on this research area "Multi-Agent System for Distributed Manufacturing Scheduling with Genetic Algorithms and Tabu Search." She has published three books and over than 150 papers in international scientific conference proceedings, journals, and book chapters.

JOSÉ LUÍS REIS received the Ph.D. degree in technologies and information systems from the University of Minho and specializes in management and administration. He is currently an Assistant Professor with the University of Maia, where he is also the Director of the Department of Business Administration and coordinates the Master's Program in Digital Transformation. Additionally, he is also an Invited Adjunct Teacher with the School of Accounting and Administration, Polytechnic Institute of Porto. He is also an Integrated Researcher with the Laboratory of Artificial Intelligence and Computer Science (LIACC), University of Porto.

JOSÉ PAULO MARQUES DOS SANTOS was born in Porto, Portugal, in 1964. He received the bachelor's degree in chemical engineering from the University of Coimbra, Portugal, in 1989, and the Ph.D. degree in management, variant marketing from the Technical University of Lisbon, Portugal, in 2011. In 2005, he worked in the industry, both for companies and as an Entrepreneur in the field of technical fabrics. He returned to academia, in 2005, firstly as a Trainee Assistant with Maia Higher Education Institute, until 2006, where he is also an Assistant Professor, until 2011, as an Assistant Professor, until 2021, with the University Institute of Maia, and as an Associate Professor with the University of Maia, Maia, Portugal, where he has been the Vice-Rector, since 2021. Since 2013, he has been an External Mobility Researcher with the Unit of Experimental Biology, Faculty of Medicine, University of Porto, Portugal, where he has been an Integrated Researcher with LIACC—Laboratory of Artificial Intelligence and Computer Science, since 2023. He has been publishing in marketing, neuroscience, and machine learning journals and conferences. His research interests include these three areas, passing by employing neuroscientific techniques to study consumers, and machine learning methods to model consumers' brain signals. He is a full member of Portuguese Engineering Council and a member of the Association for NeuroPsychoEconomics, EMAC—European Marketing Academy, and NMSBA—NeuroMarketing Science and Business Association.

DANIEL ALVES DE OLIVEIRA was born in Santa Maria da Feira, in 1982. He is currently the Head of Innovation & Research with E-goi (Portuguese Platform for Multi & Omnichannel Marketing Automation) and also the Manager of the Artificial Intelligence Research and Development Team. He has been a Teacher, a Trainer, and an International Lecturer in management, marketing and communication for over 15 years. He start his academic career as a Lecturer with the University of Maia. In parallel, he began his professional career as a Marketing Director of a professional soccer club. During his professional career, he has managed projects and businesses (nationally and internationally) worth more than 100M.