INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Exploração da Multimodalidade e da Computação Afetiva para Melhoria de Estratégias de Marketing

INÊS CÉSAR DE OLIVEIRA
Setembro de 2024

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

[Exploring Multimodality and Affective Computing for Enhanced Marketing Strategies]

Inês César de Oliveira

A dissertation submitted in fulfillment of the requirements for the degree of Master of Science, Specialisation Area of Information and Knowledge Systems

Supervisor: Dr. Fátima Rodrigues
Co-Supervisor: Dr. Ivo Pereira

Porto, September 15, 2024

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

The research design has been achieved following legitimate procedures with reachable information of scientific contributions correctly reviewed and informed consent from all the participants. The contextualization of all information included obeys the methodology chosen, ensuring accurate literature premises.

I declare having collected data with the previous consent of all participants throughout the timestamp of the research and successfully applied an anonymity and security operation, to prevent any unauthorized access and identification of any individuals.

I have elected a data analytic method committed to a transparent approach, making available the documentation for scrutiny and replication. As such, any used third-party model or framework has been previously required and provided consent for research study purposes only.

I have not plagiarised or applied any form of undue use of information or falsification of results along the process leading to its elaboration. The reports have been unveiled with honesty and virtue, without selective or manipulated recordings.

I also declare disclosure of myself and the researchers involved in any potential conflict of interest.

Therefore the work presented in this document is original and authored by me, having not previously been used for any other end.

I further declare that I have fully acknowledged the Code of Ethical Conduct of P.PORTO .

ISEP, Porto, September 15, 2024

Dedicatory

The completion of this dissertation represents the result of almost two years of great effort and dedication to my academic career and the prosperity of my professional career. Blood, sweat and tears, from the first minute to the last, which culminated in this great work that would not have been possible without the support of many people, to whom I dedicate this great and special thank you.

First of all, I would like to thank my family, both present and absent, and my boyfriend for putting up with all the efforts and battles I put myself through. In adversity, they gave me all the motivation and understanding, and in success, their presence to celebrate with me the obstacles overcome. Without their presence, especially my mother's, nothing would make sense and nothing would have been achieved! Mom, we're almost there!

I would also like to thank and acknowledge the tireless work of my ISRC research team: to Professor Fátima Rodrigues, for accepting my request for guidance, captivating and motivating me with her professionalism, dedication and experience as my supervisor; to my project supervisor, Professor Ivo Pereira, for his professionalism and friendship in helping me through all the hours (of which there were many) of supervision, for his experience, based on encouragement for research and recurring learning; to Professor Vera Miguéis, for her dedication and assiduous collaboration in all the developments, marked by sharing and collaboration throughout all the developments. I would also like to thank Professors Ana Maria Madureira and Susana Nicola, for their support and professionalism in the valuable contributions they provided, marked by their usual friendliness and care. It was a challenging journey, but one marked by many successes due to their contributions to my work.

I would like to thank ISEP, all the teachers and my fellow Master's students, who have provided me with an enriching academic journey.

To all my friends who today and always accompany me and celebrate with me every stage of this hard and challenging journey.

I couldn't end without thanking myself for not having listened to the imposter in my head, for never giving up on anything and for putting my heart and soul into every challenge, and for reconsidering taking a PhD in the near future.

Abstract

Driven by the increasing convergence of digital and physical experiences in the Marketing realm, the complexity of customer behavior has grown significantly. The emotional subtleties of these intricate interactions are becoming more difficult to fully capture using conventional unimodal techniques, particularly concentrating on evaluating textual or visual information in isolation. The urgent need for improved tools to apprehend the emotional insights of consumer preferences across multiple channels has never been more crucial than now to keep up with market competitiveness. Undertaken as part of the PHYNHANCAI project, this dissertation investigates the potential of Multimodality and Affective Computing to enhance Marketing domains. The study focuses on many modalities, including text, visual, audio, and even tabular signals, intending to uncover the affective computing contribution and provide a more holistic understanding of multimodal consumer traits. Likewise, the present work aims to examine the possible synergies between these domains, describing the benefits and addressing the inherent issues of balanced appliances. The research conducted is divided into two phases: a systematic review using PRISMA methodology, to structure the knowledge base of the domain compilation, and practical development guided by the CRISP-DM model, tailoring marketing future solutions with the review insights. The acquired results demonstrate that incorporating multimodal data leads to more accurate emotional predictions and deeper insights into consumer emotions. This dissertation also discusses ethical and legal considerations associated with multimodal AI and affective computing, providing compelling findings to improve emotional awareness in marketing strategies.

Keywords: Affective Computing, Customer Behavior, Marketing, Multimodal Artificial Intelligence, Sentiment Analysis, Systematic Review, Trustworthy AI

Resumo

Impulsionada pela crescente convergência de experiências digitais e físicas no domínio do marketing, a complexidade do comportamento dos clientes aumentou significativamente. As subtilezas emocionais destas intrincadas interações estão a tornar-se mais difíceis de captar totalmente utilizando técnicas unimodais convencionais, concentrando-se particularmente na avaliação isolada de informações textuais ou visuais. A necessidade urgente de ferramentas melhoradas para apreender as percepções emocionais das preferências dos consumidores através de múltiplos canais nunca foi tão crucial como agora para acompanhar a competitividade do mercado. Realizada como parte do projeto PHYNHANCAI, esta dissertação investiga o potencial da Multimodalidade e da Computação Afectiva para melhorar os domínios do Marketing. O estudo centra-se em várias modalidades, incluindo texto, visual, áudio e até sinais tabulares, com a intenção de descobrir a contribuição da computação afectiva e fornecer uma compreensão mais holística das caraterísticas multimodais do consumidor. Do mesmo modo, o presente trabalho visa examinar as possíveis sinergias entre estes domínios, descrevendo os benefícios e abordando as questões inerentes aos aparelhos equilibrados. A investigação conduzida está dividida em duas fases: uma revisão sistemática utilizando a metodologia PRISMA, para estruturar a base de conhecimento da compilação de domínios, e o desenvolvimento prático orientado pelo modelo CRISP-DM, adaptando as futuras soluções de marketing com os insights da revisão. Os resultados obtidos demonstram que a incorporação de dados multimodais conduz a previsões emocionais mais exactas e a conhecimentos mais profundos sobre as emoções dos consumidores. Esta dissertação também discute considerações éticas e legais associadas à IA multimodal e à computação afectiva, fornecendo descobertas convincentes para melhorar a consciência emocional nas estratégias de marketing.

Acknowledgement

This research was funded by national funds through the FCT — Fundação para a Ciência e Tecnologia, within PHYNHANCAI project: http://doi.org/10.54499/2022.01303.PTDC

Contents

List of Figures

List of Tables

List of Symbols

r_{ij}	normalized value of the *i-th* alternative for the *j-th* criterion
x_{ij}	original value of the *i-th* alternative for the *j-th* criterion
n	number of matrix entries
v_{ij}	weighted normalized value of the *i-th* alternative for the *j-th* criterion
w_j	importance weight value of the *j-th* criterion
S_i^+	the positive ideal separation distance for the *i-th* alternative
S_i^-	the negative ideal solution separation distance for the *i-th* alternative
C_i	relative closeness result for *i-th* alternative

List of Acronyms

AFEW-VA	Acted Facial Expressions in The Wild- Valence and Arousal.
AI	Artificial Intelligence.
ANN	Artifical Neural Networks.
AR	Augmented Reality.
Att-2D-CNN	Attention-based 2D Convolutional Neural Network.
Att-Bi-LSTM	Attention-based Bi-directional Long Short-Term Memory.
AUs	Action Units.
BERT	Bidirectional Encoder Representations from Transformers.
Bi-GRU	Bidirectional Gated Recurrent Unit.
Bi-LSTM	Bidirectional Long Short-Term Memory.
BoVW	Bag of Visual Words.
C3D	3D Convolutional Network.
CBAN	Crossmodal Bipolar Attention Network.
CGM	Computer Graphics Metafiles.
CMJRT	Cross-Modal Joint Representation Transformer.
CNN	Convolutional Neural Network.
CRISP-DM	Cross Industry Standard Process for Data Mining.
CTR	Click Through Rate.
DGNN	Deep Graph Neural Network.
DL	Deep Learning.
DT	Decision Tree.
E-T	Eye-Tracking.
ECD	E-mail Campaign Dataset.
EEG	Electroencephalogram.
EU	European Union.
FBT	Fourier-Bessel Transform.
FCT	Fundação para a Ciência e a Tecnologia.
FFT	Fast Fourier Transformer.
GB	Gradient Boosting.

GDPR	General Data Protection Regulation.
GPS	Global Positioning System.
IEMOCAP	Interactive Emotional Dyadic Motion Capture.
INESCTEC	Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência.
ISEP	Instituto Superior de Engenharia do Porto.
ISRC	Interdisciplinary Studies Research Center.
K-NN	K-nearest neighbors.
LDA	Latent Dirichlet Allocation.
LIWC	Linguistic Inquiry and Word Count.
LSA	Latent Semantic Analysis.
LSTM	Long Short-Term Memory.
MFCC	Mel-Frequency Cepstral Coefficients.
MIMN	Multi-Interactive Memory Network.
ML	Machine Learning.
MLBP	Multi-Scale Local Binary Patterns.
MOSEI	Multimodal Opinion Sentiment and Emotion Intensity.
MOSI	Multimodal Corpus of Sentiment Intensity and Subjectivity.
MOUD	Multimodal Opinion Utterances Dataset.
MTM	Multi-Task Model.
NLTK	Natural Language Toolkit.
PCRNN	Principal Component Regression Neural Network.
PHYNHANCAI	Enhancing Phygital Marketing through Multimodal Artificial Intelligence.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
PS-Mixer	Polar and Strength Vector Mixer Model.
QMN	Quantum-Like Multimodal Network.
ResNet	Residual Networks.
REWOA-DBN	Random Evolutionary Whale Optimization Algorithm-Deep Belief Network.
RF	Random Forest.
RNN	Recurrent Neural Network.
SDK	Software Development Kit.
SURF	Speeded Up Robust Features.
SVC	Support Vector Classifier.
SVM	Support Vector Machine.

TF-IDF	Term Frequency-Inverse Document Frequency.
TGANN	Text-Guided Attention Neural Network.
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solutions.
UMAIA	Universidade da Maia.
VAD	Voice Activity Detection.
VGGNet	Very Deep Convolutional Networks.
ViT	Visual Transformers.
VR	Virtual Reality.
Wi-Fi	Wireless Fidelity.

Chapter 1

Introduction

The present chapter pretends to provide the main theme of this dissertation, highlighting the relevance of incorporating a multimodal perspective analyzing affective computing to enhance Marketing strategies. The research on this topic is being carried out as part of the Enhancing Phygital Marketing through Multimodal Artificial Intelligence (PHYNHANCAI) project ([1]), proposed by the Interdisciplinary Studies Research Center (ISRC) based at Instituto Superior de Engenharia do Porto (ISEP) and is motivated by the exploration of customized solutions for the phygital (physical + digital) world of Marketing strategies. To this end, it is proposed that emotional tasks complement the analysis of different measurable modalities and ensure a better understanding of current public trends. The introductory chapter also presents the methodology chosen for the literature review and the structure of the rest of the document.

1.1 Project Motivation

Understanding and interpreting human interactions between the physical and digital worlds designates one of the final goals for brands to validate their success in Marketing campaigns. 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 [2, 3]. The exploitation of these improvements takes into account various factors that cut across Marketing domains, requiring the choice of tools to make the process flexible to the inherent context [4–6]. The use of Artificial Intelligence (AI) offers a wide range of alternatives [7], highlighting consumer patterns as the basic knowledge expected of their preferences and future behavior [8–10]. These developments are popularly associated with Marketing strategies due to the simplicity of implementing such a complex model, which can leverage dependencies and patterns not captured by humans [11]. The effectiveness of its applications is one of the topics currently under investigation [3, 5, 6, 10, 11], with the inclusion of emotional cues on the data under study being one of the most explored [3, 8, 9, 12–14]. As stated by Cesar et al., Gandhi et al, and Tomar et. al [13, 15–17], 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. Instead, the application of the former to the latter makes it possible to balance the complexity of the performance of multimodal models and can be applied at different times of assessment as a complement to multimodality [13, 18]. This adoption guarantees the concatenation of more information, in different formats, complementing the knowledge generated about the consumer, and their preferences, among other data [9, 13]. It is common to find methods that scrutinize emotions and feelings on opinions or other

interactions made with and by consumers ([7, 8, 13]). Expressive forms of communication differ considerably from physical to digital reality, with different approaches to sentimental analysis [5]. Research into these stimuli covers the entire journey of engagement with customers, capturing their preferences and how these can be incorporated into the next [6, 11, 14]. Prediction, attention, and classification mechanisms are some of the algorithms employed to cover the analysis of verbal and non-verbal data [12, 14].

Affective Computing is a multidisciplinary field that integrates research focused on developing systems capable of understanding, triggering, or predicting human emotions [19–24]. These technology-enhanced systems play a crucial role in various business sectors, demonstrating their flexibility and significance in personalizing user experiences [20, 24]. Understanding human emotions is a complex task [21, 22], yet it has paved the way for advancements that bring technology closer to consumers' cognitive abilities, enabling systems to better respond to their needs [23]. Recent innovations in Affective Computing, particularly through the study of multimodal human interactions, have generated valuable scientific insights into behavioral and psychological signals in consumer behavior [20–22]. As sentiment analysis and emotion recognition become more prevalent across various fields [20, 23, 24], it is essential to ensure that these systems operate in secure and reliable environments, safeguarded from data misuse or manipulation [25–29]. Ethical and regulatory considerations are increasingly important, with governments worldwide taking steps to address these concerns. For example, nations like the European Union have initiated efforts to create a global regulatory framework for AI, aimed at protecting user privacy, security, and authentication while preventing potential failures [25–29]. To ensure reliable applications, developers and creators of intelligent systems must prioritize the stability and security of these technologies [25, 26], building customer trust and ensuring confidentiality in AI-driven environments [27–29].

Considering these factors, the project motivates the development of new alternatives, promoting an in-depth exploration of Multimodal AI and Affective Computing within the context of a practical, phygital solution aimed at enhancing Marketing strategies. This approach necessitates not only understanding the current advancements in Marketing but also ensuring that marketers possess the necessary expertise to effectively integrate multiple modalities in AI applications [18, 30]. In addition to technological and strategic considerations, it is crucial to address ethical and legal frameworks, distinguishing between low- and high-risk practices to guide system design and ensure responsible, secure, and impactful implementation. This combination of technical, legal, and practical perspectives will foster the development of a comprehensive solution that bridges AI and Marketing in a real-world context.

1.2 Problem Statement

The steady evolution of Digital Marketing has led marketers to constantly strive for success in the applied strategies. Dealing with many external conditions, one of the most erratic is understanding consumer behaviors. To guarantee a weighted resolution, the recognition of customer emotions has been a major technique that focuses attention on studying previous interactions from collected data. Traditional approaches to this versatile element are achieved with unimodal computation, mostly performed with the natural processing of textual samples. However, the evaluation of the accuracy prediction of these systems to factual shreds of evidence generates mistrust in using a single modality for such predictions, proposing multimodality as an innovative way to close this knowledge gap. The heterogeneity enriches the learning process dynamic due to the multiple core challenges of fusing

information and aligning them to prioritize pertinence and suppress redundancy. Other constraints focus on the availability of data with reliable compositions, adequate computational resources, and background multimodal cognition.

This study seeks to analyze the current state of the art of multimodal approaches to the investigation of sentiment expressed by consumers in different social interactions. The blend of Affective Computing and Multimodal AI enables a better contextualization of the next Marketing strategies application, dealing with data captured to improve the transversal impact of Marketing in the phygital world. Gathering this information, this dissertation aims to enhance the development of an optimized solution considering both real-case and synthetic scenarios to study the affective computing in behavioral patterns of customers, by analyzing the insights in the multimodal experiences with different brands.

1.3 Objectives

The multimodal perspective of AI makes it possible to discover more in-depth knowledge from previously discarded sources. The heterogeneity of the verbal - text, images, audio, video - and non-verbal - Eye-Tracking (E-T), Electroencephalogram (EEG), Global Positioning System (GPS), Wireless Fidelity (Wi-Fi) - evidence creates a premise for innovation in identifying the success or failure of advertisement strategies. The information gained from studying the different experiences between customers and companies allows us to validate current trends and estimate the future direction of commerce, both physical and digital.

To complement this knowledge, this project aims to investigate the future of Marketing, in its various areas of application, through Multimodal AI. To this end, it is proposed to carry out a systematic review of the current state of the art in Multimodality to reflect on scientific advances in the field. It is also suggested that the possible combination of Affective Computing tasks, like sentiment analysis and emotion recognition, expressed and captured physically and digitally, can be investigated. The study of different alternatives allows for the implementation of AI models for recognizing emotions from heterogeneous customer modalities.

1.4 Methodology

The research methodology that covers this project has been customized to the various outlines specified in the problem statement and the objectives it covers. Thus, the use of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and Cross Industry Standard Process for Data Mining (CRISP-DM) methodologies allows for the independent documentation of all the phases, both theoretical and practical, that make up the comprehensive study of the knowledge generated on the subject.

According to [31], the benefits of applying the PRISMA methodology are further integrated into the preparation of the various procedures that make up the systematic review of scientific contributions. Its application is in line with the transparency and reproducibility requirements demanded in the validation of the permissions and results obtained. Its use highlights the relevance of this contribution compared to the work carried out by [12, 13, 17], which explores the various applications of sentiment analysis, stating the same type of literature review without the total and concrete practice of the processes belonging to PRISMA.

The presentation and discussion of the results obtained by applying the PRISMA methodology make it possible to introduce the use of the CRISP-DM methodology for documenting the phases and the correct use of the process model. This allows the life cycle of the proposed practical developments to be guided, as noted by [32], who advocates the use of the CRISP-DM model as a method of managing the reporting and evaluation of the artifacts generated from one phase to the next.

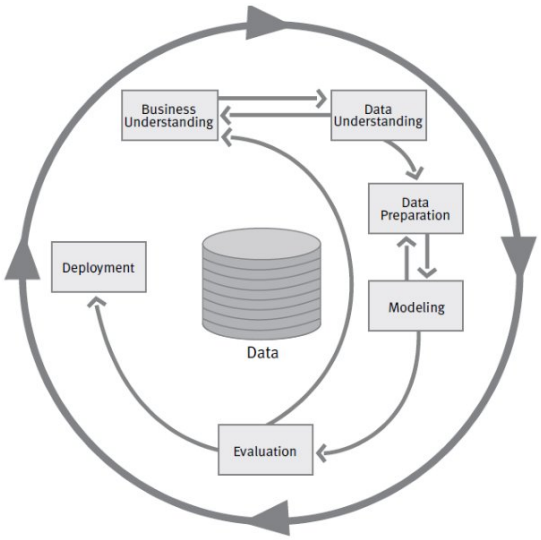

Figure 1.1: CRISP-DM Process Model, based on [32]

Figure 1.1 exemplifies the cyclical performance of the different phases that make up the model, starting with understanding the business and ending with the implementation phase. For the specific case of this project, the first phase is embedded in the main conclusions obtained from the systematic review and also structures prerequisites to be considered for understanding the data. As the problem is defined by multimodality, the understanding of the data is governed by the perception of the different representations and structuring. It is also possible to find the influence of the theoretical conclusions formed for the remaining phases of the model, relying on the practice of techniques that maintain the multimodal essence during the preparation of the data and its modeling. In these phases, continuous validation and subsequent documentation for the evaluation phase of state-of-the-art models investigated is also expected. The final implementation includes all the procedures carried out, highlighting the need to repeat one or more procedures to maximize the robustness of the proposed solution.

1.5 Structure

The six fundamental chapters that make up this dissertation each tackle a different aspect of the developed work on Multimodal AI and Affective Computing in Marketing within phygital environments.

The Introduction, in Chapter 1, introduces the motivation for the research by addressing the challenges faced in the real world in understanding consumer emotional inherent state across the multiple marketing interaction channels. The section also highlights the problem statement and presents the objectives of enhancing emotional prediction by using multimodal

data. The methodology is also documented, raising the benefits of using a two-phased approach for literature review and practical development.

In Chapter 2, the State of the Art provides the conducted systematic review of the combination between Multimodality, Affective Computing tasks, and Marketing domains. The structure follows the pre-required steps to perform the literature background using PRISMA, followed by a discussion of the results delivered according to the research questions aligned with the problem statement and the main objectives.

Chapter 3 details the Methodology for the practical design, starting with the broader research context. Then, the description of the custom application of CRISP-DM is guided by the custom conditions of this dissertation, structuring the successive procedures of data collection, preprocessing, and experimental phases. The section also covers the overview of the employed datasets, the experimental design for evaluating the multimodal models for predicting emotional labels, and the ethical measures taken during the whole process.

The Data Analysis and Experimental Design are the main topics of Chapter 4. It gives a summary of the datasets that were utilized and describes their features as well as the techniques used to extract, analyze, transform, and represent the data. The experimental design, the generation and integration of multimodal information, and the fusion strategies used in the sentiment prediction tasks are also covered in this chapter.

In Chapter 5, the Experimental Results are explained, evaluating the performance of different models applied to both datasets. The results are discussed comparatively, with a focus on how multimodal inputs improved affective computing accuracy, comparing a high-level dataset with a modulated dataset to emotional approaches. The chapter also explores the broader implications of these findings for the included domains, particularly in describing the following strategies to promote them in enhancing consumer insights and engagement strategies in the phygital domain.

Finally, Conclusions are pinpointed in Chapter 7, summarizing the key findings and discussing the limitations encountered during the study. The chapter also suggests future research directions, according to the overall challenges faced during the whole development.

Chapter 2

State of the Art

This chapter explores the current State of the Art and the contextualization of the matter encompassed by this proposal. For this purpose, the research methodology adopted to study multimodality in the sentiment analysis for Marketing is explained. Following this, a description is given of the results obtained and a detailed discussion of the common and distinct topics, creating a knowledge base to support future decisions in the development of a solution.

2.1 Research Methodology

The conducted systematic review employs the PRISMA methodology to accurately consolidate the most recent scientific contributions to multimodal Marketing on customer sentiment. This systematic review stands out from the previously analyzed [12, 13, 17] by using this methodology to seek answers to the application of this approach to the dynamics that make up the work of marketers. In addition, a contextualization of works related to practical applications is intended, analyzing details such as the modalities covered, their integration, and the ethical and regulatory considerations highlighted in the inherent challenges and opportunities.

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 [31], the PRISMA methodology 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. This only happens through the development and consideration of inclusion and exclusion criteria that positively or negatively validate the article to be considered in the review. Adjusting the practice of these instructions guarantees continuous filtering until the best resources are obtained.

2.1.1 Research Questions

According to the project objectives, the exploration of multimodality alternatives for implementing Affective Computing tasks is suggested to highlight the introduction of new Marketing strategies, enhancing both digital and physical realities. To understand the fulfillment of each domain in new proposed systems, a series of research questions are formulated as guidelines for the PRISMA methodology application.

Table 2.1: 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 these 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 Affective Computing to AI models 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 2.1 identifies the sequence of research questions that will be answered in the results of retrieved publications section.

2.1.2 Scientific Repositories

The selection of scientific repositories represents one of the first decisions to conduct a systematic review. As noticed in the work of [31], the 2020 update on PRISMA methodology 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 2.2. The parameters in each database are defined to mitigate major differences between them and will be explored later in this section.

Table 2.2: 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/`

2.1.3 Search Terms

The definition of search terms is carried out to give contextual linkage between the project objectives and the planned investigation. Providing a combination of keywords that correlate with the specified domains orientates the retrieval of articles, offering a more reliable spectrum of scientific contributions [31].

Table 2.3: Search Terms

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

As presented in 2.3, 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 Multimodality, Affective Computing and Marketing were expressed so that the concatenation of all would be included in a single string query, as portrayed in Table 2.4.

Table 2.4: Research Query

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

2.1.4 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 2.5 and 2.6 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 [31].

Table 2.5: 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 Affective Computing
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 2.6: 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 other 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

2.1.5 Publications Extraction

The extraction of publications took place under the assumptions of the PRISMA methodology, defined by three phases: *Identification*, *Screening*, and *Included*. Figure 2.1 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, defined as the last 3 years (between 2020 and 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 2.5 from the previous section. However, it was not possible to avoid specifying certain commands inherent to the interface that each virtual database requires. For SR1, the "ALL" command was defined for all the terms in the query previously presented, making it possible to choose free access. 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, with these details alone, the total result was more than 10,000 references, the first 500 references were selected. In SR3, the query was introduced without the need to introduce filtering commands to the documents, presenting the option of selecting the

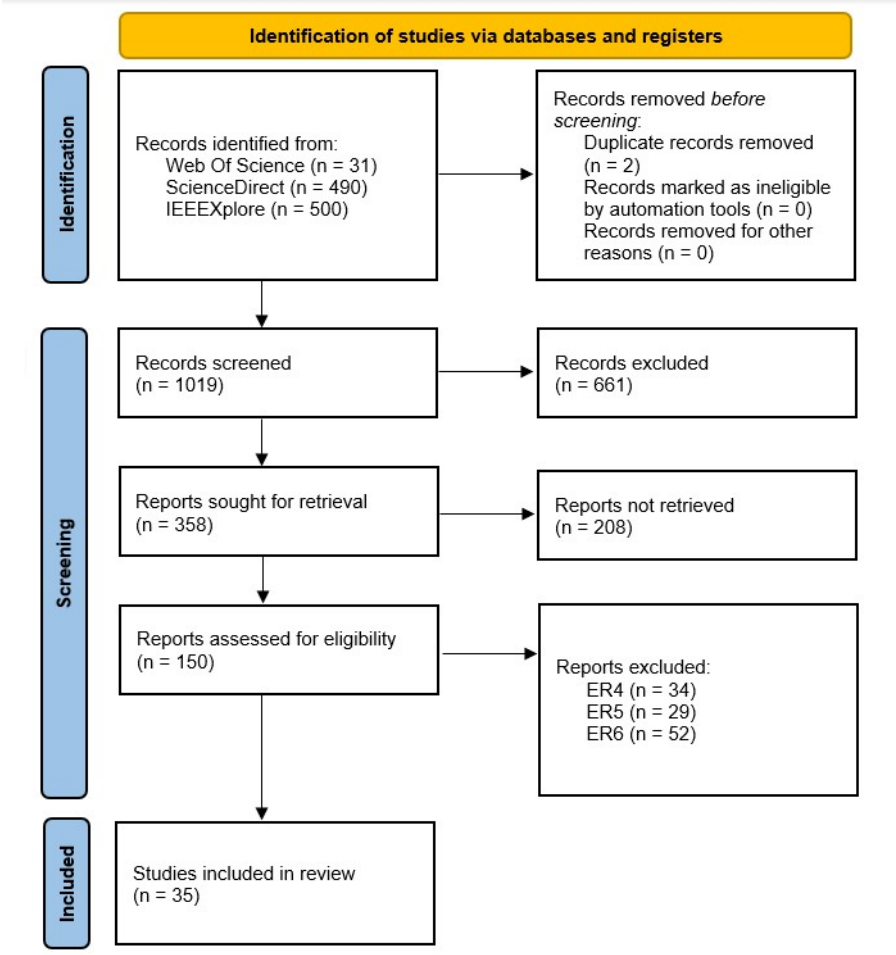

Figure 2.1: PRISMA Diagram with the retrieved articles, according to [31]

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 1020 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 involves scanning all the abstracts and excluding those that do not meet the criterion of relevance to the objectives of the review. 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, 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 2.2 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.

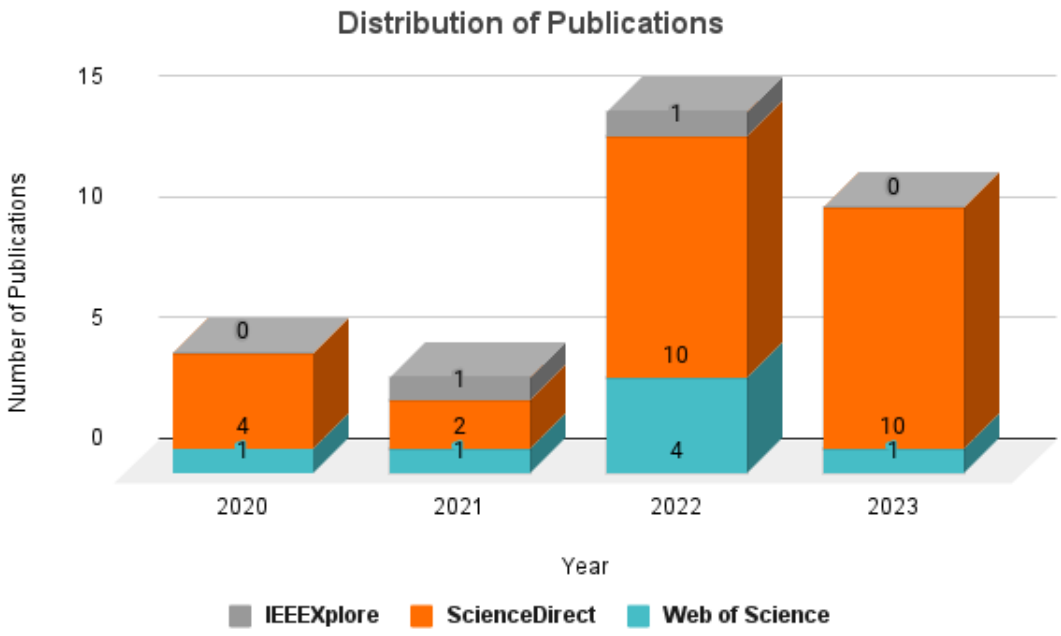

Figure 2.2: Distribution of Retrieved Publications by Research Year and by Repository

2.2 Results

This section details the obtained articles by proceeding with the presented methodology. It aims to document all the publications retrieved and the pertinence of their selection to reply properly to the investigation inquiries.

2.2.1 Multimodal Sentiment Analysis in Marketing: Current State of Research

Marketing efforts currently face challenges and adversities due to fluctuations in customer behavioral trends [33–36]. With the increasingly active parallel presence of consumers in the digital world [33, 37–41], it is necessary to foresee how complementarity with the real world can benefit brands [42–47]. The multimodal approach is therefore at the forefront of the most knowledge-intensive alternatives [35, 44–46, 48–51].

The search returned 30 articles with relevant information that allows us to describe the current state of Marketing from a multimodal perspective [36, 38, 52–55], with different categories of scientific contributions. In addition to practical applications that contextualize this innovation in the real world, to be reviewed and analyzed in more detail later, articles were also obtained in which the prosperity of new forms of communication is achieved through the use of AI [33, 34]. Taking advantage of current technology, and combining it with

the capture and transmission of information in different formats [35, 44–47, 49–51, 55], makes it possible to monitor consumer perceptions of the stimuli launched 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 [33, 34, 39, 43, 48]. The emotional weighting of these interactions linked to the multiple interpretations of data raises the uniqueness of new options for marketeers [41, 44–47, 53].

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 [34] conceives a framework capable of establishing collaboration between AI and Human Intelligence in Marketing, focusing on possible system optimizations. This relationship makes it possible to bring together the strengths of both intelligences, giving solutions mechanical, analytical, intuitive, and emotional robustness. [56] 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.

2.2.2 Multimodal Sentiment Analysis Performance Across Marketing Domains

After gaining an in-depth knowledge of the type of modalities used and how they are integrated into the emotional study in the area of Marketing, the next step is to figure out how the various 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 [33, 35, 37, 39–41, 43, 53] reviewing products [35, 36, 48, 57] and also predicting elements such as the prices of services [58] and the popularity given by consumers [38, 52, 54]. The importance of studying direct and indirect interactions between companies and customers, evaluating the dialogue established [50, 51, 55] and the public opinion generated is also highlighted [44–46, 49, 59–62]. 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 [33, 37, 39, 40, 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 [33, 37, 39–42]. Although it is referred to as an area of Marketing application under investigation [33, 37, 39–43], 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 [33, 37, 40, 43], but ensures more flexible participation by the parties involved, allowing them to create content that can be integrated into subsequent strategy updates [37, 39, 42]. As a result, quickly obtaining opinions by associating emotional expressions [37, 39, 42] gives way to understanding current customer trends [33, 37, 39, 40, 42] by analyzing the evolution of likes, comments, and shares using likehood models [33, 37, 40, 43].

The employment of the multimodal approach in service or product review analysis has a positive effect on consumer engagement [35, 36, 48, 49], leading to an increase in E-commerce sales [35, 36, 49]. This is what the authors in [36] explore, 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 [35, 48, 49], reducing customer uncertainty at the time of purchase [35, 36, 49]. However, there are problems associated with this increase, such as the increase in returns of products bought online [35, 36] and the discrepancy in the data obtained from questionnaires aimed at regular customers [35, 48, 49]. The heterogeneity of the characteristics obtained by product evaluation [35, 36, 48] interconnects their fusion as a way of overcoming these situations [35, 36, 48, 49], highlighting the research work considered by the authors [35, 48] to be preliminary on these techniques. They see the use of emotion stimulated in these reviews, some through tone and frequency of voice [35, 49], as a way of predicting future audience behavior [36, 48].

It is also quite typical to use multimodality as a way of granting recommendation systems greater coverage of the contextual content of each problem [38, 53, 58]. Combining the disparities of each variable ensures better ways of developing advertising [38, 52, 58], recognizing not only the content as the basis of all the dynamics but also the foreseen target audience [38, 52, 54]. The use of social networks, capable of supporting the recognition and development of new content [52, 53, 58], allowing it to extend its prediction to other outcomes but made up of the union of multimodal representations [38, 52–54, 58]. The performance of predictive models expects the flexibility of applying them to a phygital reality [54, 58].

The use of sentiment analysis in dialog interactions [50, 51, 55] provides a new angle of perception that is difficult to capture in modality with text in images. Developments in interpreting conversations with feelings [50, 51, 55] are still a growing topic of study and practice due to the associated complexity [50, 55]. This type of data structure validates interactions between humans as well as robots or other types of responsive systems [55], 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 [44–46, 49, 59–62]. Social networks once again play a leading role as a platform where opinions are obtained in different formats [44–46, 49, 59, 60] 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 [44–46, 49, 59–62].

2.2.3 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 [33, 35–46, 48–55, 57–64]. 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.

.

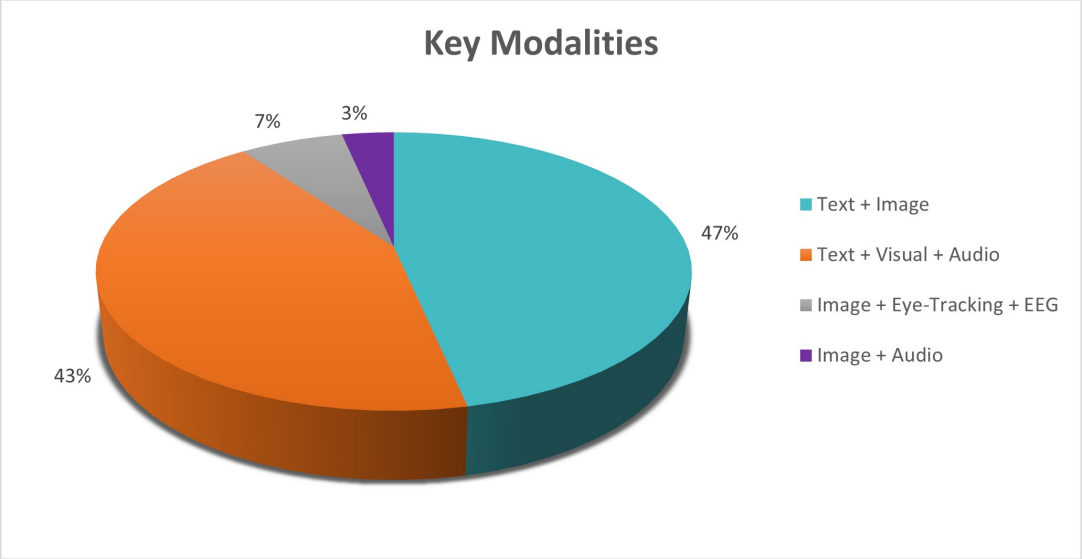

Figure 2.3: Modalities used in retrieved articles

By analyzing the content documented by the different articles, we can see that 7 categories reflect the combination of modalities, expressed in Figure 2.3. The circular graph confirms the dominance of the fusion between text and images present in 47% of the articles [33, 36–41, 43, 46, 47, 50, 52, 59], 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 [35, 44, 45, 49, 51, 53–55, 60–63]. In third place, with 7% of uses, auditory evidence is replaced by audiovisual resources, guaranteeing greater diversity accompanied by high complexity [48, 64]. 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 by the model highlight the integral progress in generating knowledge. Other categories, such as the inclusion of series capturing E-T and EEG, complementary with images, or even the study of the impact of audio only with images or with text, are the least mentioned, but they are novel because of their computation with AI. In this way, and 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 2.4 graphically presents the datasets the scientific community adopts on emotional categorization in human interaction activities in Marketing. Datasets such as Taobao [35, 52], Multimodal Corpus of Sentiment Intensity and Subjectivity (MOSI) [45, 49, 51, 53, 54, 57, 60–62], Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) [45, 49, 53, 54, 61, 62], and Multimodal Opinion Utterances Dataset (MOUD) [44, 57, 60] aggregate data samples referring to reviews and feedback expressed by consumers on

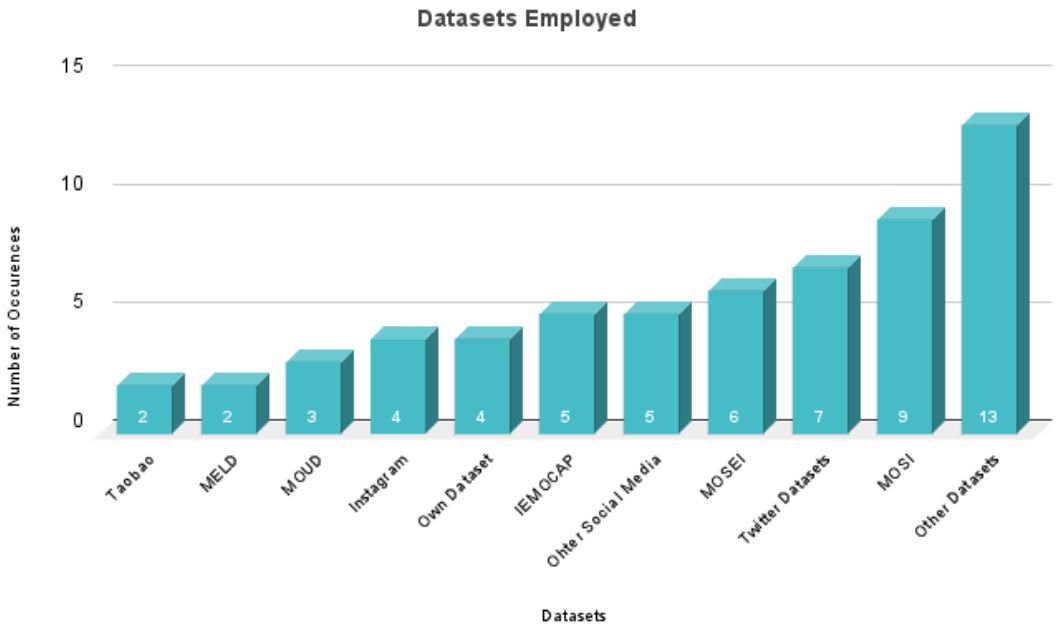

Figure 2.4: Datasets employed in retrieved articles

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 [50, 55], Interactive Emotional Dyadic Motion Capture (IEMOCAP) [44, 50, 53, 55, 63], and "Other Datasets" [36, 37, 39, 42, 43, 49, 52, 58] 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 [36, 38, 39, 42, 43, 47, 57], such as the speed of Instagram [33, 39, 40, 43] and Twitter [37–39, 41, 59]. 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 [46, 48, 53, 64] 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.

2.2.4 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 2.3, 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 text, images, audio, or video samples, it is possible to perceive the associated valence ([33, 35–39, 41–43, 45, 46, 48, 51, 52, 54, 57–61, 63]), the emotional weight expressed in sentiments ([49, 53, 64]), or both ([40, 44, 50, 55, 62]).

Tables 2.7, 2.8 and 2.9 present a resume of each technique specified for integrating the different modalities in each article.

Based on the review undertaken, text and image emerge as the two most commonly used features in building multimodal perspectives for AI integration. While not all studies predict sentiment using Affective Computing with AI models—such as Gu et al. and Gandhi et al. [33, 42], where polarity is calculated through likes and comments to test multimodality's impact—there are notable similarities in the underlying procedures. In textual data extraction and preprocessing, techniques like GloVe [41, 43, 46, 50] for word embeddings and the Google Cloud Vision API [33, 38, 39, 42] for textual extraction (either directly or via OCR on images) are frequently used. Additionally, custom versions of the Bidirectional Encoder Representations from Transformers (BERT) model [37, 59] are popular for text processing. Other techniques, like Linguistic Inquiry and Word Count (LIWC) [33], Fourier-Bessel Transform (FBT) filters extending Latent Dirichlet Allocation (LDA) [52], and neural networks such as Convolutional Neural Network (CNN) [50] and Bidirectional Long Short-Term Memory (Bi-LSTM)(Bi-LSTM) networks for contextual analysis [46], are also utilized. Additionally, attention mechanisms [59] and Bidirectional Gated Recurrent Unit (Bi-GRU) [41] are incorporated for handling semantic details, while techniques like Latent Semantic Analysis (LSA) and Term Frequency-Inverse Document Frequency (TF-IDF) [36] enhance preprocessing capabilities. Libraries such as the Natural Language Toolkit (NLTK) [38, 42, 43], String and regular expressions (re) [42], and web-scraping tools like Selenium and Beautiful Soup [40] also play a key role. Image captioning techniques, such as Computer Graphics Metafiles (CGM) [59] or Bag of Visual Words (BoVW) [39], are similarly essential in multimodal processing.

Table 2.7: Summary of Integrations With Textual and Visual Modalities, according to ([18])

Ref.	Textual	Visual	Fusion Approach	Models	Baselines
[33]	LIWC, GVAPI	GVAPI	Hybrid	Likelihood	Poisson Model, AIC, BIC
[36]	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	-
[50]	GloVe, Density Matrix, LSTM, NLTK	SIFT, k-means, LSTM	Hybrid	QMN	CNN, FMF, DSEF, MDL, CRNN, h-LSTM, QMSA, DM-CNN, DM-QIMF

Continued on next page

Table 2.7: (continued)

Ref.	Textual Preprocessing	Visual Preprocessing	Fusion Approach	Models	Baselines
[46]	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
[41]	GloVe, MHSA, Bi-GRU	ResNet, Capsule, Attention Mechanism	Attention	EF-Net	ATAE-LSTM, IAN, MemNet, MGAN, Res-MemNet, Res-IAN, Res-MGAN, ESFAN
[52]	FBT+, LDA, LSTM, MLP, MCB+FFT	VGGNet, LSTM, MCB+FFT	Joint/Bilinear	TGANN	LR, SVM, CAN, LightGBM
[37]	BERT	ViT	Attention	CBAN-Add, CBAN-Dot	MVAN, MultiSentiNet, CoMN, FENet, MMHFM, VilBERT, LXMBERT, 2D-Intra-Attention+RoBERTa, VGG + CNN, Relation-Attention, Transformer-Attention
[59]	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

Continued on next page

Table 2.7: (continued)

Ref.	Textual Preprocessing	Visual Preprocessing	Fusion Approach	Models	Baselines
[39]	GoogleLens, SentiCircle + ConvNet, GloVe, VADER	GoogleLens, BoVW, LBP, k-means, SVM	Hybrid	HCConvNet-SVM	SVM, Naive Bayesian, KNN, Gradient Boosting
[40]	langdetect, SentEMO	Scraper, SentEMO	-	RoBERTa	-

The Google Cloud Vision API is commonly used for both textual and visual feature extraction [33, 38, 39, 42], competing with models like Very Deep Convolutional Networks (VGGNet) [52, 59], Visual Transformers (ViT) [37], CNN [43, 50], Long Short-Term Memory (LSTM) [52], and Residual Networks (ResNet) [43, 46]. These models may or may not integrate attention mechanisms [46]. Additionally, other methods like scraper software [40], Speeded Up Robust Features (SURF), and Multi-Scale Local Binary Patterns (MLBP) [36] further contribute to the rich diversity of techniques applied to both modalities. This variety allows for the extraction of features that are highly contextualized to the specific problem, making the fusion process unique in each case. The use of graph-based models, such as in Jouyandeh et al.'s study [38], enables dynamic updates of node weights as new data is added, employing Greedy and Random algorithms to compare elements for polarity prediction. Attention mechanisms linked to neural networks, especially in fusion and sentiment analysis models, have been noted as a positive trend in research by Qian et al., Cheung et al., Lopes et al., and Gu et al. [37, 41, 43, 52], enhancing model performance. However, the design of the models is often influenced by how each modality's data is used for predicting sentiment, either individually [38, 42] or in combination during [59] or after [36, 37, 40, 41, 43, 46, 50, 52] the fusion process.

The fusion approach in multimodal integration plays a pivotal role in leveraging complementary information from text and visual modalities. Hybrid approaches, as seen in the works of Gu et al., Mehbodniya et al., Gandhi et al., Zhang et al., Lopes et al., and Kumar et al. [33, 36, 39, 42, 43, 50], combine various techniques to extract features from both modalities before feeding them into the model. These approaches capitalize on the synergy between textual and visual information, thus improving predictive performance. Intermediate fusion methods, like those employed by Zhou et al. [46], process textual and visual data independently before merging their representations, allowing for specialized preprocessing and capturing more nuanced relationships between the modalities. Meanwhile, attention-based fusion strategies, such as those used by Gu et al. and Cheung et al. [37, 41], dynamically adjust the weight given to textual and visual features during inference, prioritizing the most relevant information. This adaptive approach is particularly effective when the importance of modalities varies across instances, leading to improved prediction accuracy. Qian et al. [52] and Xiao et al. [59] take a similar approach with joint fusion, where modalities are first processed independently, and then their features are combined, taking into account the specific performance requirements of each context.

In terms of model selection and comparison, several architectures have been proposed to handle multimodal data and harness the synergy between text and image features. Models such as Text-Guided Attention Neural Network (TGANN) [52], Crossmodal Bipolar Attention

Network (CBAN) [37], and CoolNet [59] incorporate advanced neural network components with attention mechanisms to capture complex relationships within and between modalities. These models use specialized activation functions, optimization algorithms, and regularization techniques to enhance learning and generalization. Additionally, ensemble methods like Random Evolutionary Whale Optimization Algorithm-Deep Belief Network (REWOA-DBN) [36], Quantum-Like Multimodal Network (QMN) [50], and Hybrid CNN-Support Vector Machine (SVM) (HCConvNet-SVM) [39] aggregate the outputs of several base models to improve overall performance. State-of-the-art models such as MultiSentiNet [37] and Multi-Interactive Memory Network (MIMN) [46] serve as benchmarks, providing a reference point for evaluating new approaches and assessing performance improvements.

In the work presented by Chen et al. [58], both audio and visual modalities form the foundation for developing an emotion recognition system designed for price analysis on an e-commerce platform. The input consists of video samples of customer behavior and opinions, which are decomposed into granular audio and visual data. For audio processing, the SoundNet CNN, noted for its effective learning abilities, is employed to extract features such as Mel-Frequency Cepstral Coefficients (MFCC) [58]. These audio features are further transformed using the Fast Fourier Transformer (FFT) algorithm. On the visual side, RGB-based frames are extracted from the videos using DenseNet CNN and VGGNet networks. These neural networks recognize sentiment at both the singular image and plural scene levels [58]. The classifiers used for these modalities include an LSTM with a batch normalization layer, which combines the modalities via model fusion based on weight voting.

Table 2.8: Summary of Integrations with Textual, Visual and Audio Modalities, according to [18]

Ref.	Textual	Audio	Visual	Fusion Approach	Models	Baselines
[44]	Google Translator, word2vec, CNN, LSTM	C3D, 3DCNN, ConvLSTM, FC-LSTM	openSMILE, SVM	Hybrid	3DCLS	CNN, RNN, C3D
[45]	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

Continued on next page

Table 2.8: (continued)

Ref.	Textual	Audio	Visual	Fusion Approach	Models	Baselines
[53]	PERT, Mean Pooling	ViT, Mean Pooling	Spleeter, wav2vec, wav2vec2D	Attention	MTM	MFH, MFB, MLB, Joint-encoding, MulT, EF-MTM w/o res, LF-MTM w/o res, MTM w/o res
[35]	ELECTRA, SnowNLP, word2vec	ResNet	openSMILE	Attention	Open Tran-former	DNN, Audio Transformer, wav2vec, HuBERT, Bi-GRU, Bi-LSTM
[49]	Normalization, Recode, CNN, word2vec, Bi-LSTM	3DCNN	openSMILE, Min-Max Normaliza-tion	Attention	DSAGCN	CNN, bc-LSTM, CMN, DialogueRNN, DialogueGCN, AGHMN
[63]	PTWE, Conv-BiGRU, Bi-RNN, Deep CNN	DPTM	librosa	Attention	AMSAER	BLSTM, HMM, DBN, DCGAN, CBP, CMN, ML-SER
[60]	textual LSTM	openFace2, Facet, vi-sual LSTM	COVAREP, openS-MILE, acoustic LSTM	Attention	Tree Adaptive Frame-work	CHF, CAF, MAG, SMM, SEP, CTP, ITP
[61]	Bi-GRU, GloVe, CNN	Co-Attention Matrix, Facets, 3D-CNN	COVAREP, openSMILE	Hybrid	-	MFN, Graph-MFN, DCCA
[54]	Textual LSTM	Visual LSTM	Audio LSTM	Attention	-	CAT-LSTM, Simple LSTM, Hierarchi-cal RNN, Contextual RNN

Continued on next page

Table 2.8: (continued)

Ref.	Textual	Audio	Visual	Fusion Approach	Models	Baselines
[51]	TF-IDF	openCV, dib, Mediapipe	openSMILE, pocketsphinx, librosa	Late	Tri-Feature Fusion	Hfusion
[57]	LDA, IBM Watson, GloVe, CNN	CERT	OpenEar	Attention	BLSTM(MA)	MFN, MARN, GME-LSTM(A), TFN, BC-LSTM, EF-LSTM, Bi-LSTM
[55]	GloVe, BERT	OpenFace2	COVAREP	Hybrid	AOBERT	C-MFN, TFN, LMF, MISA, Graph-MFN, MTMM-ES, TBJE-2, TBJE-3
[62]	GloVe	Facet	COVAREP	Late	CMJRT	EF-LSTM, LF-LSTM, TFN, LMF, MulT, MISA, Self-MM
[58]	-	DenseNet, VGGCNN	SoundNet CNN, FFT	Late	Q-learning algorithm	SWM, RF, LSTM

The fusion approach integrates textual, visual, and audio modalities for emotion analysis, utilizing a variety of strategies. Hybrid fusion, seen in models like Polar and Strength Vector Mixer Model (PS-Mixer) [45] and OpenTransformer [35], combines features from different modalities before feeding them into the model, leveraging the complementary information from each modality. Attention mechanisms are crucial in dynamically weighting the contributions of each modality during inference. For instance, DSAGCN [49] and Multi-Task Model (MTM) [53] apply attention mechanisms to selectively emphasize relevant features from the various modalities, allowing for adaptive integration based on task-specific requirements. In contrast, late fusion approaches, as demonstrated by the Cross-Modal Joint Representation Transformer (CMJRT) [62], delay the integration of information from each modality until after the individual modality predictions have been made. This approach enables the model to process data from each modality independently before combining the results, allowing it to capture more complex inter-modal interactions.

The models applied in these studies utilize a range of advanced neural network architectures and algorithms to process multimodal data. Architectures such as CNNs [44, 49], LSTMs [35, 49, 51, 54, 55, 58, 62, 63], Bi-GRUs [35], and Transformer variants [35, 45, 49, 53, 55, 62, 63] are commonly employed to capture complex relationships both within and between

the modalities. Models like OpenTransformer [35] use Transformer architectures with attention mechanisms to process and fuse multimodal inputs, while DSAGCN [49] leverages CNNs and LSTMs for visual and auditory feature extraction, respectively. Traditional architectures like Recurrent Neural Network (RNN)s and 3D Convolutional Network (C3D) are also frequently used as baselines against which the performance of these advanced approaches is measured. By comparing their proposed models to these benchmarks, researchers are able to quantify the improvements in handling multimodal data that arise from fusion strategies and advanced architectures. This comparison provides valuable insights into the effectiveness of each approach for multimodal emotion analysis tasks.

Table 2.9: Summary of Integrations With Visual, EEG and E-T Modalities, according to [18]

Ref.	EEG	E-T	Visual	Fusion Approach	Models	Baselines
[64]	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
[48]	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

The methods for integrating data from visual, EEG, and E-T modalities across different applications are detailed in the Fusion Approach column. A common technique used is hybrid fusion, which merges features from multiple modalities at various stages of model development. For example, Wu et al. [64] implemented a hybrid fusion strategy that combines visual inputs with EEG data through the use of pattern recognition algorithms and EEG signal processing techniques. Similarly, Zhu et al. [48] employed a hybrid fusion approach by utilizing fuzzy algorithms to integrate eye-tracking data with EEG signals. This approach allows the model to capture complex patterns and relationships by leveraging complementary information from different modalities.

In terms of neural network architectures and signal processing techniques, the selected models demonstrate a range of approaches for interpreting data from EEG, E-T, and visual modalities. Wu et al. [64] utilized a Deep Graph Neural Network (DGNN) to simultaneously process EEG data and visual patterns, efficiently capturing temporal and spatial relationships within the EEG data. Similarly, Zhu et al. [48] applied an Attention-based 2D Convolutional Neural Network (Att-2D-CNN) to analyze EEG and E-T data. These models improve

performance in multimodal analysis tasks by integrating attention mechanisms to focus on relevant features and patterns within the input data.

To evaluate the effectiveness of their proposed fusion methods and model architectures, both studies use conventional machine learning techniques and well-known neural network architectures as baseline models. In Wu et al. [64], baseline models such as Artifical Neural Networks (ANN)s, SqueezeNet, GoogleNet, ResNet, DarkNet, Inception, and Inception-ResNet are used to compare against the performance of the proposed DGNN architecture. These baseline models provide a reference for assessing the ability of the fusion techniques to analyze multimodal data effectively. Likewise, Zhu et al. [48] compared their Att-2D-CNN model against baseline models like SVM, Random Forest (RF), Principal Component Regression Neural Network (PCRNN) and Attention-based Bi-directional Long Short-Term Memory (Att-Bi-LSTM)s, highlighting the advantages of their proposed method.

2.2.5 Ethical and Regulatory Considerations

Of the 38 articles returned, [28] and [29] 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 ([28]). Even though Marketing is one of the central investigation areas, mentioned in various sections in [28] and [29], 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 ([28, 29]).

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 ([29]). 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 [28] and [29] 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 (EU), with a set of articles regulating the use of the intelligence generated by these innovations ([28]). Both [28] and [29] 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 ([28]) 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 ([28, 29]). 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 ([29]). [28] and [29] 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 ([29]). 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 ([28, 29]).

In [28], 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 [28]. 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 [28, 29]. The author in [28] translates 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, [28] 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.

[29], on the other hand, provides 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, [29] emphasizes 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 [28, 29]. 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 [28, 29].

Although the approaches to the ethical and regulatory considerations of AI taken by [28] and [29] are disparate, 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 [28], 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 [29], 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 [28, 29]. 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. [29] also mentions 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 [28], 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 [29], 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 [29], the ethical and regulatory considerations for Marketing encounter several challenges to be overcome [28]. 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.

2.2.6 Challenges and Opportunities for Potential Future Applications

The exploration of the different applications of Multimodality and Affective Computing for Marketing imperatively needs to recognize a series of existing challenges and possible promising opportunities for the future of this approach [28, 29, 56, 65]. As it is a new way of working, it acquires a diversity of interpretations considering the real case in question [38, 56]. 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 [28, 29, 65]. The design of a reliable system relies on the choice and concatenation of data that is significant to the problem in question, creating external factors that influence non-controllable changes [28, 56]. In this way, data integration has an impact on the rest of the system's performance and is one of the major problems with its constitution [37, 43, 53, 60]. Not knowing the nature of the samples made available to the models compromises the guarantee of their effectiveness due to the lack of data quality [42, 48, 52, 65]. 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 [36, 47, 52]. Thus, this type of obstacle can lead to arguments that justify documenting the inefficiency of multimodal adoption compared to the traditional adoption of just one modality [33]. It is crucial to document all the details that could affect the results obtained, in line with the ethical and regulatory recommendations mentioned above [28, 29, 65].

In addition to the obstacles mentioned, which may be part of the subsequent consequences, it is also important to mention the lack of transparency in the perception of the work carried out by the model [43, 53, 61, 64, 65]. 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 [28, 29, 43, 65]. 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 [38, 43, 53, 64]. 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 fickle and difficult to moderate, its scalability compromises its application in real-time, without ethical and structural considerations being ruled out [38, 51, 54, 65]. 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 [28, 29]. 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 [37, 65]. The need to legalize certain procedures and restrict others places the application of AI as an unwise practice without first formulating all these considerations [56, 65].

To overcome these obstacles, the spotlight of today is focused on the incentive to formulate new promises for the future use of Multimodality and Affective Computing [42, 43, 65]. 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 [33, 42, 65]. 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 [38, 52, 54]. Knowing current trends leads to a flexible application to the requirements of each sector of activity [38, 51], 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 [39, 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 [28, 29]. 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 [51].

Chapter 3

Methodology

The following chapter describes the methodology employed to operate multimodality in affective computing within marketing contexts. Given the intricate diversity in data modalities and the urge for accuracy and interpretability, this study adopts the CRISP-DM method baseline to personalize the work phases and align with the overarching research objectives. The systematic nature of the multimodal procedure study ensures a rigorous data exploration, enabling the navigation through the technical challenges while considering ethical deliberations, which are paramount to the analysis of the consumer sentiment implicit and explicit feedback. The chapter documents each stage throughout these concerns while systematically contributing to the broader understanding of Affective Computing in Phygital Marketing.

3.1 Research Context

The versatility of multiple modalities of data supplies a deeper, more complex insight into consumer behaviors, attracting many marketers to use it as an essential mechanism. Traditional marketing analytics, built to focus on discrete data points like purchase history or web browsing, uncovers many limitations to capture the intertwined customer preferences and interests, resulting in an implicit and unique emotional valence response. Tailoring marketing strategies resonating with consumers on an emotional level, although challenging, plays a vital role in elevating market competitiveness with more robust predictions of future acquisitions, and brand loyalty reinforcement. This is particularly important when it comes to omnichannel initiatives since more than ever, clients demand coherence and consistency across the physical and digital interactions with brands.

Resonating all of these concerns, the present research is conceived as one of the main objectives of the PHYNHANCAI project, held by the ISRC at ISEP and funded by Fundação para a Ciência e a Tecnologia (FCT). Entitled Enhancing Phygital Marketing through Multimodal Artificial Intelligence, the project explores the improvement of Marketing strategies regarding phygital opportunities in omnichannel experiences, highlighted by the implementation of Virtual Reality (VR) and Augmented Reality (AR) technology to analyze and compute the affective evolution of the Customer Experience. PHYNHANCAI proposes the use of Multimodal AI, focusing on Affective Computing as one of the main tasks to investigate the future of new developments in Marketing, improved by the uniqueness of each consumer's preferences.

The PHYNHANCAI project is a very cooperative initiative that unites the knowledge and assets of many noteworthy partner organizations, each of which makes a distinct contribution to the achievement of the project goals. As the project leader, ISRC is crucial in organizing

the study, offering theoretical and technical guidance through the vast experience in data science and engineering of the selected academic members, both professors and students, complying with strict scientific standards and new and creative engineered solutions.

The further associated partners bring critical and invaluable support to the project and to the main task, ensuring both scientific and industry knowledge on marketing domains and applications. E-goi, a Portuguese company focused on improving E-commerce through a marketing automation platform to enhance every business omnichannel interaction, provides a deeper understanding of market trends and consumer behavior. Fulfilling the need to be aligned with real-world communication environments, E-goi's cutting-edge technology enables this project to develop a solid foundation for testing and validating the innovative data-driven strategies to improve marketing practices with Affective Computing, accessing the multimodality inherent in customer interactions.

Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESCTEC), a private non-profit research association with more than 30 years of pioneering developments on scientific and technological research initiatives, is also one of the PHYNHANCAI project partners. Their expertise significantly contributes to integrating essential and methodological rigor and advanced technologies to gather and transfer valuable knowledge into the study environment. As an institution that operates symbiotically with the academic and business worlds, their dynamical vision helps leverage the project journey enhancing the quality and impact of findings.

Universidade da Maia (UMAIA), best known as an academic excellence institution in marketing education, furnishes a critical perspective on the challenging paths of the domain. This associative work conducts a practical and relevant contextualization of Marketing impact, ensuring highly valuable use of theoretical tools and frameworks. The involvement of UMAIA guarantees academic robustness and rigor as well as shaping the path for the next generation of marketers.

These collaborators come together to establish an outstanding consortium that blends academic rigor with business observations, generating a synergy that greatly increases the project's potential to impact academic research as well as the marketing sector. The PHYNHANCAI project has become efficient in investigating and improving the phygital marketing approaches using Multimodal AI and Affective Computing thanks to the varied contributions from ISEP, E-goi, INESCTEC, and UMAIA.

3.2 Methodology Overview

To systematically address the complexities of multimodality enhanced in the usage of Affective Computing within Marketing communication channels, a customized methodology based on the CRISP-DM process model was designed. Implementing a cyclical method that allows a continuous refinement of the data collected and generated grants flexibility and iterative research on the impact of multiple modalities to analyze the dynamic preferences of the target public. This methodology was personalized to highlight the particularities of the real problem context, exploring different approaches to reducing complexity among the heterogeneity of behaviors and data.

The CRISP-DM model is organized into six phases, providing a business and data understanding as baselines of the following decisions. Business Understanding focuses on translating specific data mining goals and determining factor criteria to scope the main objectives to

fulfill. At the same time, Data Understanding involves the collection, exploration, and assessment of data samples that facilitate insights about the available data to sample and explore their granularities, defining a modality unit. In this way, the next phase entitled Data Preparation, thrives to proceed with cleaning, transforming, and structuring the variables for multimodal feature engineering. This process focuses on performing a collection of data techniques to analyze the heterogeneity and patterns of abstraction for knowledge alignment among them, one of the requirements that anticipate the fusion approach in the Modeling phase. It focuses mainly on the application of different models and algorithms to the prepared datasets, organizing a report of unlocked knowledge resourcing on the diversity of modalities. Matching not only with the model's performance metrics but also with the research objectives, the Evaluation assessment leverages the predictions with the obtained results and validates the framework value. Finally, Deployment consists of possible tests monitored on real-world environment cases.

Taking into consideration all the procedures presented, a new methodology is proposed to satisfy the need to explore different approaches to multimodality to uncover different affective cues on marketing analytic metrics and heterogeneous variables that compose advertising interactions. However, the configuration follows a deep concern about taking advantage of different compounds of data without reflecting ethical and privacy issues for the third-party intervenients.

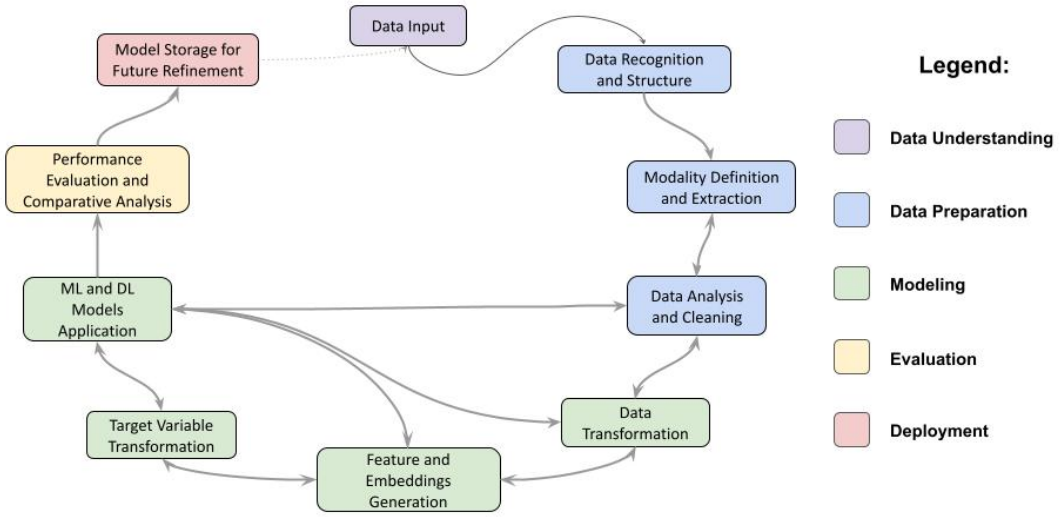

Figure 3.1: Customized Methodology Diagram based on CRISP-DM process model

Figure 3.1 provides an overview of the methodology employed in the research, following a customized CRISP-DM model approach to structure the implementation of multimodality and affective computing, throughout the phases of data analysis to the deployment. The alignment with CRISP-DM model phases is described in the figures' legend, distinguished by the different colors and annotations. The methodology begins with the Data Input stage, where the first examination is conducted to get more insights into the available collection and to correspond the data understanding process with the business context already acquired.

Next, the Data Recognition and Structure represent the bridge between the data's first inquiry and the first procedures to effectively start the Data Preparation phase. Recognizing different data characteristics and structuring them into a reproducible format are some of the achievements granted in this phase, simplifying the complexity of using raw data in affective computing tasks. Consequently, Modality Definition and Extraction, a complementary step from the previous one, defines and selects those variables that could be conveyed into modalities. The extraction of these data points is obtained by dealing with data granularities and displaying them in the most suitable format. After outlining the modalities that compose the customer interactions, Data Analysis and Cleaning perform the last techniques in the Data Preparation phase, by dealing with missing data and outliers. It also refines the previously mentioned steps by organizing each variable to be joined up with similar ones, forming the modalities, or with similar appended charges for the outcomes. Then, after preparing all the available information, the Modeling phase starts by transforming data into suitable designs to capture relevant connections among different features. The Data Transformation considers the transition of tabular data, increasing capabilities and consistency on the additional information that enhances the multimodality. Feature and Embeddings Generation is a complement phase of the previous one, by performing analogous techniques but for modalities that were defined earlier. The heterogeneity and the abstraction for each single modality captured are some of the main preserved attributes to generate more robust embeddings, passing the crucial details into the learning process of models. But before proceeding to the application of Machine Learning (ML) and Deep Learning (DL), the target variable must be remodeled as well to baseline the emotional perspective on customer behavior metrics. After managing all the data, the models' application is designed and set in an experimental environment to evaluate the performance and the results integrity, by running a comparative analysis on different use cases and specifications. Once the Evaluation phase is completed, the Deployment is granted by storing the model and preparing for future refinements.

3.3 Data Collection and Preparation

The process of collecting and preparing the data is broken down into different stages, impacted by the perception of the current state of the market in terms of the frequency of multiple modalities and the incitement of emotional signals in customer interactions. Considering the main objectives of this research, it is necessary to resort to choosing and preparing the data to be applied in the model shown in Figure 3.1. Given the search for new and possible improvements in the evaluation of marketing interactions with the target audience, it was necessary to resort to different use cases that could reveal different evidence of consumer feedback, both implicit and explicit. In this way, the work developed is based on the study of two datasets, captured in different contexts of interaction with customers and with different levels of associated detail, which aim to comprehensively explore the different perspectives of the use of multimodality and affective computing. The datasets were collected considering two cases of application: a dataset inserted in the real context of the marketing solutions market, collaboratively obtained with the partnership of E-goi, and another public dataset for scientific use, proven as a baseline for studying multimodality on tasks in the field of affective computing. The differences inherent in the characteristics and structure of the datasets make it possible to explore the benefits and implications of use cases that translate the extremes of sentiment analysis through multimodality. However, it is in the translation of details, granularities and scientifically validated annotations that the enrichment of possible

marketing integrations lies, descriptively evaluating the impact of the rigor and detail of the modalities on the performance of the entire prototype.

The first use case consists of data from E-goi, a company at the forefront of omnichannel marketing. It collects and stores data daily relating to its interventions that take place directly or indirectly on the platform developed. These, in turn, are carried out on different communication channels to understand customer behavior and track physical and digital presence through the attractiveness of campaigns. In line with the various models and aspects of marketing applied by E-goi, this information is collected employing metrics on the sending of campaigns by different means, such as the frequency with which they are carried out and their reception by the target audience. In this way, communication via e-mail was considered the preferred channel, usually composed of textual and visual evidence strategically arranged to attract the user's attention and trigger instant interactivity with its recurrent sending. In addition to this, other data is generated to provide supplemental temporal and tabular information to contextualize campaign schedules.

The distributed arrangement of this information in different files made it difficult to perceive and prepare the constituent modalities at a high level, requiring an elementary reorganization of each one into an accessible and easy-to-manipulate data structure. Using the Beautiful Soup Python library, visual and textual modalities were isolated from the HTML that made up the entire e-mail layout, considering logos and visual hyperlinks ignored. On the other hand, the text features are complemented by the concatenation of every textual evidence, linking the subject, snippet, and HTML text to define the textual modality to a single representation. The relevance of the time variables is also identified, as well as the results of the campaign performance reported to improve the final purpose of exploring this dataset. Taking into account the nuances associated with marketing, the application of metrics that help to analyze strategic insights was developed, making the transformation of the objective variable encompass the frequency of sends and opens, the number of clicks, and redirects to other associated functionalities. However, while it is possible to infer the implicit nature of consumer feedback from the presence or absence of clicks, it is also difficult to explore clear affective evidence of the stimulus involved. These limitations associated with the concrete application of all the associated domains have led to the need to select the second dataset as a way of overcoming all the constraints present in this one, while also seeking a series of guidelines for future improvement.

Technique for Order of Preference by Similarity to Ideal Solutions (TOPSIS) is a multi-criteria decision-making method that evaluates alternatives based on their distances from the positive and negative ideal solutions. By comparing a set of alternatives for the best affective computing exploratory dataset to use, the definition of some criteria was conducted to evaluate each one accordingly. To each criterion, a score was given based on the importance of the project objectives to calculate the geometric distance between each approach, deciding which one is the most appropriate. Given the transparency nature of this decision method with full access to the operations of each step, TOPSIS was employed as an automated tool to choose the most suitable dataset, supporting the research objectives of optimizing multimodality and affective computing for marketing.

Table 3.1 documents the criteria determined for this particular decision-making task, supplemented with the importance weights for the final decision. The method implementation assumes 5 criteria that specify the relevant configurations for the study of multimodality in sentiment analysis for marketing. Firstly, Data Quality regards the accurate and reliable nature of the data preparation to enhance the affective assignment of analyzing sentiment

Table 3.1: Criteria Importance Weights

Criteria	Importance Weight
Data Quality	0.25
Dataset Size	0.15
Relevance to Affective Computing	0.35
Modality Diversity	0.15
Availability and Accessibility	0.10

Table 3.2: Decision Matrix of the MOUD, MOSI, and MOSEI datasets

Dataset	Data Quality	Dataset Size	Relevance to Affective Computing	Modality Diversity	Availability and Accessibility
AFEW-VA	90	600	9	7	8
IEMOCAP	80	10 000	7	7	8
MOUD	70	400	8	7	6
MOSI	80	2 199	8	8	9
MOSEI	90	23 500	9	8	8

through the different modalities. Ensuring completeness in emotional details of the defined labels, aligned with the modal inputs, significantly influences the success of the model performance and the suppression of minority scenarios that can mislead the results. The size of the dataset, similar to the previous one, also affects the robustness and generalization of the model learning process. The dataset dimension qualifies the proposed solution with more examples to generate the knowledge base, becoming vital to capture a wider variety of emotional expressions across each modality to the fused representation of all. However, it is necessary to balance the quality of data with its size, probing for improvements and validation of the research findings. To achieve this, the Relevance to Affective Computing is measured by considering the main features and objectives of the named datasets. Due to the inherent complexity, this criterion is the most weighted since datasets must build a solid baseline of detailed emotional content, attached to the modalities that support multimodality research. The same goes for Modality Diversity, which delivers a multi-dimensional perspective on the data representation and alignment, capable of enhancing the learning process and supplying different inferences on the knowledge generation. By evaluating the number of modalities considered in each dataset, a set of hypotheses can be raised to design possible signs of progress of the proposed framework. At last, the Availability and Accessibility of each collection must be simplified, engaging more conducted studies to overcome the recent limitations.

To initiate the TOPSIS method, a decision matrix must be designed with suggestive numeric estimators, translating the presence of each criterion. However, each can be distributed in different formats, given in percentage, real values, or even a score in some defined range. As presented in Table 3.2, the Data Quality is scaled as a percentage considering the overall stats of available insights. On the other hand, Dataset Size is defined by holding the number of samples disposable by each collection, demonstrating the diversity disposed by

each option. At last, the following criteria are represented as a score from one to ten values. The Relevance to Affective Computing, the most weighted aspect to satisfy, regards scaling the developed work in each dataset to sharpen the emotional prediction tasks. This criterion is achieved by studying the structure, quality, and ground truth strength to perform one or multiple affective computing assignments with the same collection. Likewise, Modality Diversity is also considered, expecting to gather different and robust modality representations and being uniformly arranged with emotional cues. Availability and Accessibility are two crucial requirements that guarantee the functionality of searching and manipulating with data variables, without compromising the dataset effectiveness.

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \tag{3.1}$$

The second step of the TOPSIS procedure is to normalize the decision matrix previously created. The criteria are converted into a unique unit range to be compared with one another. By taking the Normalization Equation 3.1, where r_{ij}, the normalized value of the *i-th* alternative for the *j-th* criterion, is obtained by calculating the quotient between the original value x_{ij} and $\sqrt{\sum_{i=1}^{n} x_{ij}^2}$, the normalization factor of n matrix entries. Table 3.3 displays the results of the application of the previous described equation.

Table 3.3: Normalized Decision Matrix of the Acted Facial Expressions in The Wild- Valence and Arousal (AFEW-VA), IEMOCAP, MOUD, MOSI and MOSEI datasets

Dataset	**Data Quality**	**Dataset Size**	**Relevance to Affective Computing**	**Modality Diversity**	**Availability and Accessibility**
AFEW-VA	0.517	0.020	0.525	0.500	0.508
IEMOCAP	0.459	0.343	0.408	0.500	0.508
MOUD	0.402	0.014	0.467	0.500	0.381
MOSI	0.459	0.075	0.467	0.571	0.571
MOSEI	0.517	0.805	0.525	0.571	0.508

The third phase of the TOPSIS method is based on a matrix multiplication with the weights of the criteria defined initially, in sequence to mathematically contextualize the problem in question. The associated weights define the importance of each criterion for integrating the relevant domains into the proposed solution, such as the choice of dataset. In this way, it is possible to generically adapt the understanding of all the factors mentioned and customize division processes that impact the flow of the rest of the work.

$$v_{ij} = r_{ij} \times w_j \tag{3.2}$$

Taking advantage of Equation 3.2, the normalized decision matrix is multiplied by the criteria weights, amplifying the impact of each point on the decision, and generating Table 3.4 to properly understand the balance created to a fair evaluation of each alternative.

Once all the computations have been applied to the matrix to assist in comparing all the criteria for the selection of the best dataset, a detailed study is carried out to gain a better

Table 3.4: Weighted Normalized Decision Matrix of the AFEW-VA, IEMOCAP, MOUD, MOSI and MOSEI datasets

Dataset	**Data Quality**	**Dataset Size**	**Relevance to Affective Computing**	**Modality Diversity**	**Availability and Accessibility**
AFEW-VA	0.129	0.003	0.184	0.075	0.051
IEMOCAP	0.115	0.051	0.143	0.075	0.051
MOUD	0.101	0.002	0.163	0.075	0.038
MOSI	0.115	0.011	0.163	0.086	0.057
MOSEI	0.129	0.121	0.184	0.086	0.051

understanding of the extreme scenarios. The concepts of positive ideal solution and negative ideal solution are designed to build up each of the extreme scenarios to be drawn up, acting as a benchmark for the final evaluation. Equations 3.3 and 3.4 define the positive and negative ideal solutions for this process stage.

$$S_i^+ = \sqrt{\sum_{j=1}^{m}(v_{ij} - v_j^+)^2} \qquad (3.3) \qquad\qquad S_i^- = \sqrt{\sum_{j=1}^{m}(v_{ij} - v_j^-)^2} \qquad (3.4)$$

The positive ideal solution S_i^+ represents the best scenario across the defined criteria. By appointing the highest value of each measure, the theoretical solution is assembled to aggregate the most favorable choice. Alternatively, the negative ideal solution S_i^- is used to understand how far a criterion is creating the worst-case scenario. Each one portrays the distance from the distribution edges, considering the value of each dataset v_{ij} and subtracting to the best or the worst evaluation, illustrated respectively by v_j^+ and v_j^-.

Table 3.5: Separation Distance Results for Each Dataset

Dataset	S_i^+	S_i^-
AFEW-VA	0.134	0.048
IEMOCAP	0.089	0.059
MOUD	0.140	0.019
MOSI	0.127	0.032
MOSEI	0.006	0.144

Table 3.5 displays the calculated quantification of how far each alternative is from the positive ideal solution and the negative ideal solution. While the positive ideal solution reflects the distance from the best scenario, the negative ideal solution performs the opposite purpose, both based on the Euclidean distance formula. Obtaining these measures prepares the final step of TOPSIS, where the closeness to both scenario alternatives is calculated to a preference score for the suitable choice.

$$C_i = \frac{S_i^-}{S_i^+ + S_i^-} \qquad (3.5)$$

The relative closeness C_i to the positive ideal solution is performed by using the Equation 3.5 to measure the dissolution of each dataset qualification. Dividing the negative alternative of each option S_i^- by the total of summing both ideal, represented by S_i^+, and negative solutions, the closeness is obtained to support the most preferable option to pick.

Table 3.6: Relative Closeness Results for Each Dataset

Dataset	**Relative Closeness (C_i)**
AFEW-VA	0.264
IEMOCAP	0.397
MOUD	0.119
MOSI	0.202
MOSEI	0.962

By applying TOPSIS to ascertain the best choice of dataset, it was proven that MOSEI is the most suitable for enriching the results and conclusions sought, disclaimed in Table 3.6. This dataset is, to date, the largest dataset for exploring sentiment analysis and recognizing the intensity of emotion in opinion videos and other instances published online. The genre balancing associated with this dataset makes it possible to study various topics where the speaker's inherent opinion is transcribed and properly annotated in terms of the associated context. The detail found on each modal element makes it possible to obtain a deeper understanding of the behavioral patterns of the speakers, providing benefits to the datasets without any kind of emotional intervention. These characteristics become noticeable upon extraction, with a pre-aligned organization and granular characterization, constituting the embeddings of the accepted modalities.

3.4 Experimental Design

The experimental design for the exploratory study of multimodal affective computing framework for marketing is conceived in its prototype form development phase integrating the two main sources of data earlier detailed. Intending to discover insights about emotional breakdowns in marketing-based experiences, the configuration of this research workflow focuses on recognizing each dataset composition, granting them the modality representation after cleaning and preparing data while considering the domain implication in each collection. The transformation of each modality conceived to create data modalities, assisting on the fusion approach for multimodal embeddings representation. The exploit of outcomes is divided into implicit and explicit feedback, due to each dataset's characteristics in translating customer emotional traits into their behavior. To achieve this, intermediary phases are purposely incorporated to dissect the overall setup success with different options.

Figure 3.2 illustrates the experimental design for this research project, integrating the two primary sources of data: the E-mail Campaign Dataset (ECD), provided by E-goi, and the MOSEI, created by the MultiComp Lab. The workflow presented demonstrates how each dataset was decomposed into modalities and the processing carried out to analyze and predict implicit and explicit feedback. Both results are dependent on the dataset concerned, due to different information that can lead to explicit emotional feedback, considering the valence of expressed sentiment, or implicit emotional feedback, conceived by inferring on

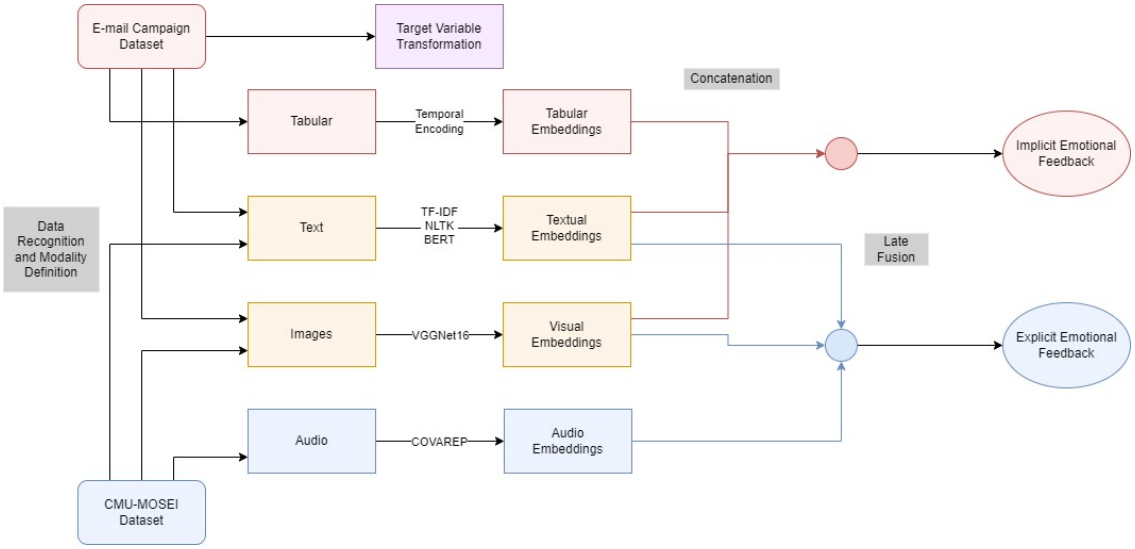

Figure 3.2: Experimental Design Workflow

overall opinions of customers and predict the emotional valence through the frequency of behaviors.

The Data Recognition and Modality Definition provided in the previous methodology diagram begins by identifying and defining the various modalities—Tabular, Text, Images, and Audio—from both datasets. From the ECD dataset, Tabular data is extracted, encoding temporal and general information to capture time-based patterns and diversity of campaigns, resulting in Tabular Embeddings. The text data is preprocessed using several techniques, including TF-IDF, NLTK, and BERT, and converted into Textual Embeddings. Visual features from the campaign images are processed through the VGGNet network and the OpenFace tool, creating Visual Embeddings. For the MOSEI dataset, similar steps are followed. Not only the visual embeddings but also textual ones will differ considering the compositions of the source text and images to produce it, leaving only one similarity: the size of the embedding. Audio embeddings are obtained, with the particularities of COVAREP, extraction of relevant acoustic features to baseline the modality presence in the model performance. The Text, Images, and Audio modalities are processed. The audio modality, in particular, uses the COVAREP feature extraction method to generate Audio Embeddings that represent various acoustic features relevant to emotional expression. The different modalities are then assisted by a late fusion approach to satisfy the integration of these knowledge sources, complementing each contribution to accurately fine-tune the model. The design is tailored to integrate and compare the emotional cues from both implicit and explicit customer interactions, providing a comprehensive analysis of emotional responses using a multimodal approach.

3.5 Ethical Considerations and Compliance

Throughout the overall methodology development, some measures were addressed considering up-to-date ethical considerations to ensure adherence to the highest standards of data protection, privacy, and transparency. Collecting, transforming, and using multimodal data requires responsible AI employment, regarding the investigation of unveiled emotional insights.

One of the first actions to take place before the beginning of data analysis is the anonymization of individuals' sensitive information. To ensure that no volunteers can be traced back from the original data subjects used in the model learning process, strict guidelines are operated according to data protection regulations like the General Data Protection Regulation (GDPR). This is a procedure that has been secured by the MOSEI developers, presenting their dataset by embedding alignment between modalities. The usage of two different datasets improves bias mitigation procedures, adding transparency and accountability for future fine-tuning optimization based on ground truth and compliance with the AI Act regulation.

Strategies for bias reduction were actively explored since machine learning models, particularly those used in sentiment analysis, are susceptible to biases. Two separate datasets, ECD and MOSEI, were used in the process, which helped to minimize the possibility of overfitting and made it possible to create models that are more broadly applicable in a variety of scenarios. By utilizing a variety of data sources that represent various facets of consumer behavior, the incorporation of several modalities - text, picture, and audio - further mitigates the danger of bias. In order to guarantee that the model's predictions are impartial and fair across demographic groups—a crucial component for practical marketing applications—the outcomes are also thoroughly examined.

The ethical considerations and compliance is especially crucial in the context of marketing, as choices made based on an emotional understanding might affect the overall interaction with how customers feel and behave. The approach places a strong emphasis on the requirement for openness in AI-powered suggestions so that users are aware of the ways in which their data is utilized to create customized marketing campaigns.

3.6 Summary

The method used to investigate multimodal sentiment analysis in marketing has been described in this chapter, with a particular emphasis on the incorporation of affective computing through the use of a tailored CRISP-DM model. The chapter began with the inquiry setting, outlining the PHYNHANCAI project and its principal stakeholders while highlighting the significance of multimodal data in contemporary marketing. A thorough description of the procedures followed in this study, including data collection, preparation, and assessment, was given in the methodology overview. The selection of the MOSEI dataset for strengthening the study was justified by discussing the usage of TOPSIS, a multi-criteria decision-making tool. This chapter also addressed the ethical issues to guarantee adherence to current data protection laws, such as the GDPR and AI Act, as well as the experimental design that facilitates the deployment of multimodal frameworks. This chapter creates a thorough framework through these procedures for the section that follows, which presents the practical development and assessment.

Chapter 4

Data Analysis and Experimental Development

This chapter summarizes the data analysis and practical development carried out in this research project to meet the empirical demands using multiple modalities for emotional analysis of marketing interactions. This section documents the practical procedures taken from data input to the implementation of Machine and Deep Learning models, and it establishes a solid baseline for validating and comparing the results obtained. Various strategies, ranging from feature engineering and embedding creation to model optimization, are also presented to demonstrate unity with the study's objectives and the suggested methodology approach.

4.1 Datasets Overview

The data analysis practices are performed on two distinct datasets, each representing a different use case and extracted from different sources according to the requirement of operating multimodal marketing experiences to analyze and obtain insights on emotional spurs. Divided into two use cases, the analytics strategies conceive the examination of each collection and aggregate the particularities to identify the most suitable approach, in agreement of disclosing a prior and independent characterization.

The dataset that makes up the first use case is provided by E-goi, which organizes its database taking into account different types of communication channels, exploiting all the customization processes provided to attract the customer to interact with. Through one of the collaborative actions carried out by the company for this project, a collection of data was assembled regarding the e-mail campaigns segment and the respective configuration variables: the delivery information, its content, and the report of the client's behavior measures. With more than 340 marketing campaigns, the present dataset has approximately 60 variables portraying the set boundaries to personalize the customer experience. Among them, the adherence of customers to the advertisement is measured by counting specific actions, like opening the e-mail or clicking on its segments. E-goi structures the forward routines by relying on these marketing analytics to yield a relationship with the target public, fitting both the dynamical trends and the available products. This effort is exhibited by detailing approaching these measures and intentionally capturing the number of times when these behaviors are unique, exemplified by the variables displaying the scenarios where the recipient only opened or clicked once.

Regardless, the connection of this dataset to the imposed domains only identifies suitable requirements for two of them, lacking congruent annotations to perform an explicit study of one or more affective computing tasks. The inherent complexity of performing an Affective

Computing approach demands a careful investigation of possible alternatives to be considered as one. However, the performance of Marketing tactics often includes the implicit report of a binary classification, whether or not it is interesting to the client, among other provisions. Claiming ethical compliance, the transformation of the implicit report measures to a multiclass target feature, similarly following the scientific processes of public datasets, is considered to be a relevant first step to align the aspects of sentiment analysis to other marketing strategies that do not involve human intermediaries expressing ground truth affective signals. The fluctuance of market kinetics highlights the importance of studying different features independently and jointly at an early stage to convert the gathered knowledge into logical target classes and more appropriate techniques.

The second use case is determined to evaluate the deep knowledge of a prepared multimodal and affective computing dataset, developed by the Multimodal Communication and Machine Learning Laboratory (MultiComp Lab) to expand the strengths of the multidisciplinary research of Multimodality and Affective Computing. The MOSEI dataset is one of the largest and richest available for the study and analysis of sentiment analysis and emotion recognition, taking advantage of more than one modality. Their composition contains more than 23 thousand video segments of extracted opinions from the online community in video blogs, lectures, and reviews, covering different topics with affective expressions from the speakers. Each segment was carefully remodeled to have manual annotations and other relevant features for emotion recognition and sentiment analysis, ideally centered on improving the affective computing tasks. The great amount of video samples provides an exponential creation of different elements for each modality breakdown and feature alignment, remodeling the robustness of model development and implementation.

Adjusting the extraction from MOSEI samples, it is possible to isolate three major modalities: textual, audio, and visual signals. Each one of them contributes deeply to the interpretation of the obtained results by consolidating complementary cues from the other ones to improve the overall understanding of the emotions and sentiments. The data capabilities for performing more than one affective task highlight the detailed process of preparing data samples and the support of training and testing many research hypotheses. For that, the CMU-MOSEI developing team pursued creating a Software Development Kit (SDK) to properly handle the data customization process and release pre-trained models to enhance the contribution of this data collection. The available modalities are delivered to their users after a preprocessing organization, represented by their embedding form alongside the allocated labels, to prevent data privacy issues for the YouTube content creators included in the dataset creation. The download of data samples is done in different frequencies, keeping the computational sequence of keys for data congruence organization. However, the alignment procedure is advised to be accomplished given the multiple segments of different annotations and labels, improving the rigorousness of the dataset ground truth.

4.2 Use Case 1: E-mail Campaigns Dataset

This section proposes the workflow of analyzing, preprocessing, and transforming the E-mail Campaigns Dataset into a valuable resource of multiple modalities to predict the implicit opinion sentiment, by evaluating the stimulus of the e-mail communication and the obtained report of interactions. Each stage pertinently discusses the options and states the benefits of the chosen techniques, while being compliant with ethical and regulatory guidelines to reorganize the aim of the real-world dataset.

4.2.1 Exploratory Data Analysis

To fully explore the available data in the E-mail Campaigns Dataset, the first step conduct a customized selection of relevant features to be considered in the experimental model training. The approach prioritizes the analysis of the variables, by focusing on emphasizing the expected outcomes. Impacted by exploring the usage of Affective Computing routines for Marketing enhancement, this procedure also investigates the alternatives that comply with assembling a new target outlook: the implicit feedback given during customer experience.

Table 4.1: E-mail Campaign Dataset Relevant Variables Extracted

Variables	Description
hash	The campaign unique identification sequence
subject	The subject of the e-mail
snippet	The snippet of the e-mail
start_date	The starting process date of e-mail campaign delivery
end_date	The ending process date of e-mail campaign delivery
scheduled_date	The scheduled date for the process of e-mail campaign delivery
content	The HTML code of the e-mail campaign content
sends	The number of e-mail campaign delivery target audience
opens	The number of times the target audience opened the e-mail campaign
clicks	The number of times the target audience clicked in the e-mail campaign segments
unsubscriptions	The number of times the target audience clicked in the unsubscription segment

For each campaign instance, the extraction of features was performed by intermediary scripts to extract and store the campaign tabular information, the content of the e-mail, and the reported results. Table 4.1 describes the selected features to constitute the final dataset composition and start the research methodology procedures of preprocessing, features detail extraction, and modality representation for the experimental setup. For each campaign, E-goi conceives a unique identification sequence in *hash* code to distinguish from the thousands of campaigns created and sent daily. As configurable e-mail sections, the e-mail's *subject* and the *snippet* are extracted as essential parts that stimulate the customer to open the received e-mail. As textual exhibits that extend and relate symbiotically, these variables were also strategically picked as suitable elements for integrating the textual modality of the dataset. Furthermore, the campaign also has distinguishable variables to express the adjustments of the date and time associated with the e-mail sent. The start and end date are associated with the arranged scheduled date, being representative of the instant that the delivery operation started and finished. These two moments are defined when the campaign is created, expressing the pretended temporal point when the campaign should be sent. Capturing these three variables unwinds the understanding of the most popular timestamps that maximize the interactions and find possible outliers that compromise the inclusion of temporal insights. The content variable stores the HTML code of the e-mail layout, to streamline the text, images, and hyperlinks fitted in the e-mail compounds, structuring the advertised offers. It is considered to be one of the most important variables, requiring the application of a series of techniques to extract new modalities and enhance

existing ones, respectively visual and textual elements. The following variables are part of the generated analysis report of the campaign assessment, identifying the relevance of considering the target audience through the number of sends, the responsive customers that opened the campaign, and the number of times it received one or more clicks. However, selecting the unsubscriptions metric, the number of clients who chose to stop receiving the e-mail campaigns, incorporates a more detailed business understanding with the appliance of Affective Computing. Commonly presented in every advertisement e-mail, the unsubscribe option also induces the count of clicks, misleading the real number of interactions. Due to that, this variable is chosen to improve the real value of users' implicit feedback, ideally to be comprised in further transformation steps.

As previously mentioned and throughout the variables selection, this dataset lacks detailing the customer experiences that could lead to investigating affective behaviors or expressions. However, the particularities of the campaign reports introduce a simple but effective way of measuring the number of exchanges made with the target audience, proportionally spectating the campaign's success. The pattern evaluation of these measures can reinforce the attempt to apply additional annotations resembling the customer's emotional preferences and supply business insights to produce a new target label. Figures 4.1, 4.2 and 4.3 illustrate the distribution of three of the mentioned metrics, which reinforce the understanding of the audience coverage range. Eventually, the need to evaluate each measure distribution consolidates the presence of possible outliers or missing data and prepares for the upcoming practices of disclosing more context-based knowledge.

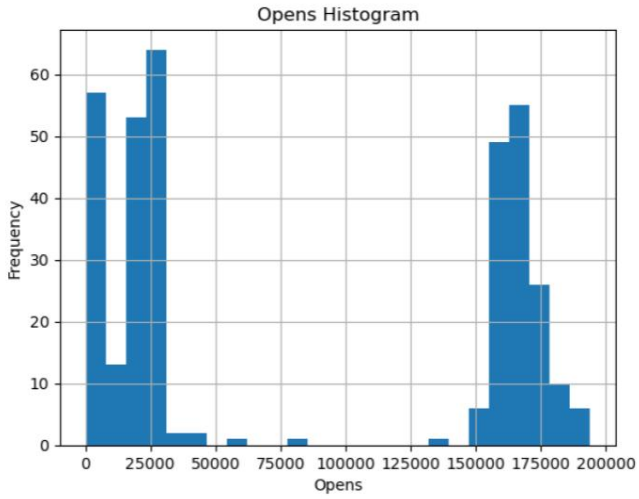

Figure 4.1: Histogram of the Opens Frequency

The open frequency histogram in Figure 4.1 struts the distribution of opens within the dataset, with the number of opens in the x-axis and the frequency of campaigns within the specific range of opens in the y-axis. By analyzing the plot, two clusters of data points are visible, granting the majority of elements being part of a bimodal distribution. The illustrated split suggests the existence of two different types of campaigns, or user segments, where the first group is defined by lower engagement with fewer opens. In contrast, the second one exhibits highly engaging campaigns with an increased open rate. Although this duality of scenarios, a few instances also represent the middle ranges of the defined x-axis, demonstrating a potential gap between the dynamical classes of interaction.

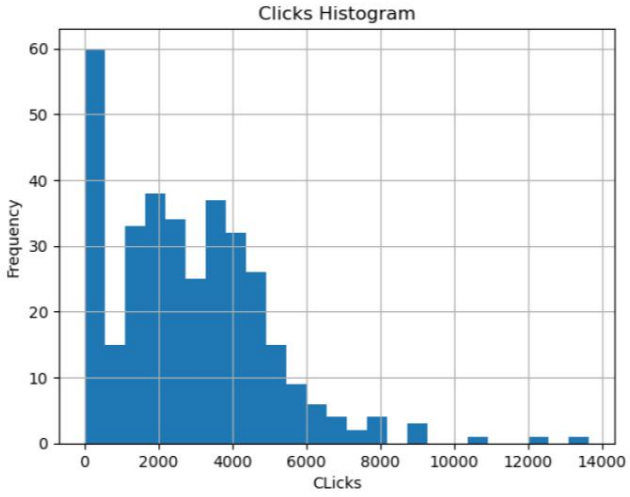

Figure 4.2: Histogram of the Clicks Frequency

As for Figure 4.2, the chart displays the disposal of clicks throughout the dataset samples. Similarly to the previous figure, the x-axis represents the number of clicks and the y-axis is the frequency of occurrences for each click range. The observations indicate most of the utterances have a click count between the first scale values, more specifically 0 and 2 000. Between these boundaries, a significant frequency of data entries has zero clicks, providing two possible issues: a large portion of unmatched campaigns to the user preferences or the presence of data discrepancies. While the first option might be the most logical one, justified by the exploratory studies conducted to predict and reverse the low engagement, the second option may also impact the results, being represented by the creation of campaign testers and experimental deliveries. As the number of clicks increases, the frequency of occurrences decreases, solidifying the hypothesis that a small percentage of campaigns achieve a significant value for evaluation metrics, like click-through rates. As seen, few instances exceed 8000 clicks while the 10000 ones are almost rare, which could also indicate the presence of data outliers. These measures, as examples of the overall report generated, translate the online interactions and respective adherence to each.

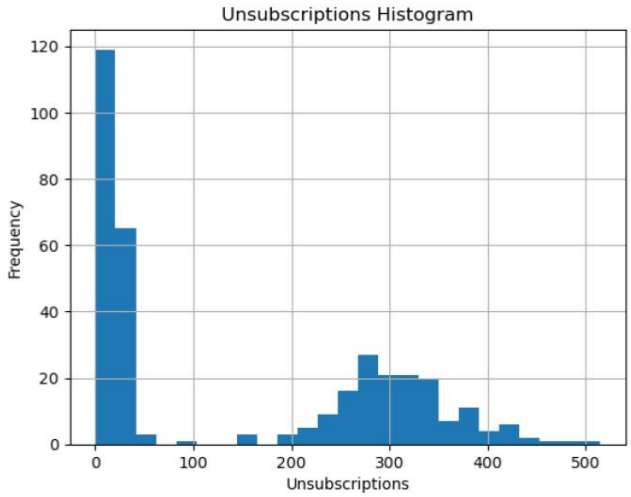

Figure 4.3: Histogram of the Unsubscriptions Frequency

Figure 4.3 provides the unsubscriptions measured across all the marketing campaigns. In ditto with the x and y-axis, the histogram notably demonstrates three major observations, which emphasize the final results. The presented chart suggests that the involved marketing campaigns are great at user retention, with few rate spike observations but noteworthy for further scrutiny. The first strong peak is near zero unsubscriptions, with over a third of the campaigns showing almost infrequently disengage to interactive campaigns. Regardless, as the number of withdraws increases, the frequency of campaigns drops abruptly but remarkably recognizing the presence of a second cluster of unsubscription. The counting range goes from 200 to 400, being highest at around 300, indicating a minority of cases where some campaigns ushered to significantly higher rates. At the same time, rare scenarios of an unsubscription range greater than 400 are shown, leading to the need to evaluate possible outliers.

Remaining compliant with the exploration of the campaigns' improvement, the dataset considers complementary information of the sequence of interaction attempts, like the temporal details of each e-mail. The Marketing strategies performed by E-goi follow a temporal alignment, studying seasonal and trending patterns, leading to a promising feature to could enhance the multimodal investigation. The choice of an appropriate time slot for engaging the contact with the clients can maximize the chances of maximizing the reported activity. At the same time, granting these temporal insights to the learning process boosts the integrity of context nuances and perfects the model outcomes.

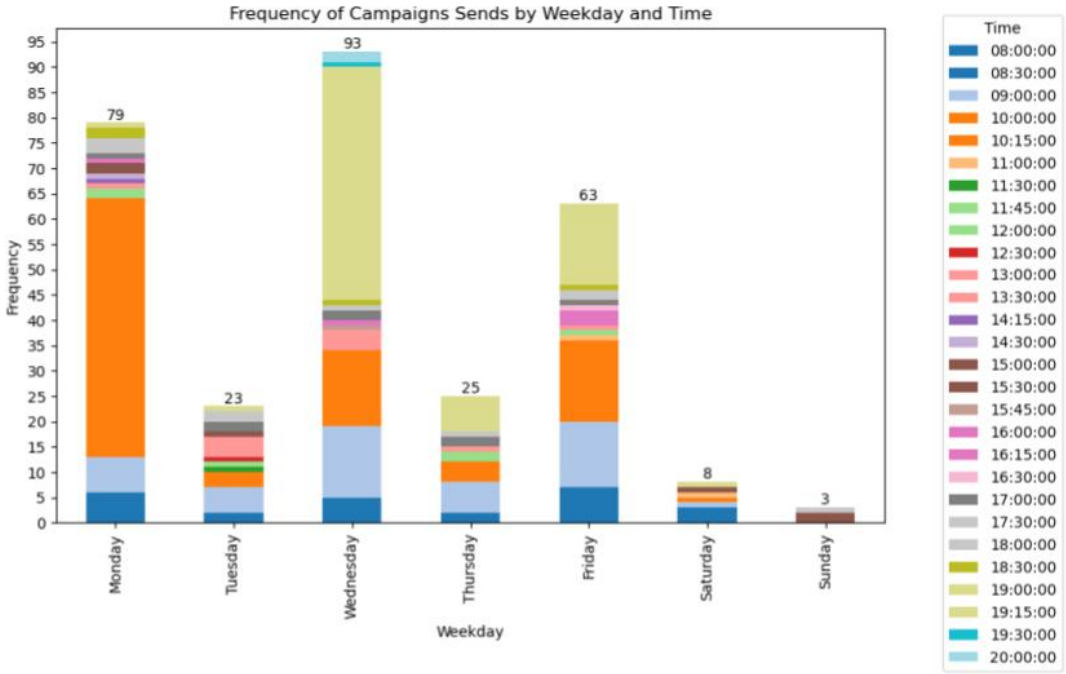

Figure 4.4: Frequency of Weekday and Time of the Campaign Send Process after Preprocessing

Figure 4.4 illustrates the generated multi-stacked chart to understand the yearly delivery of e-mail campaigns during the week, filtering by the defined weekday and time set. As the abscissa represents the days of the week and the ordinate, the frequency of sent campaigns, it is possible to visualize valuable information for marketing domain premises integration. The colored segment of the bars indicates the time of day when campaigns were sent,

described in the color legend of the plot. In each weekday, the number of sent campaigns throughout the year is displayed, suggesting an active but pondered choice of the appropriate moment to start sending the campaigns. The observations capture the most strategic actions in the beginning and middle of the week, having Monday, Wednesday, and Friday as the most popular weekdays, acknowledging the morning shift as the dominant one. The intermediate days, Tuesday and Thursday follow a similar trend of the early hours but reveal mild interactions, with an abrupt decrease during the weekend. The blended analysis of the temporal distribution implies the benefits of engaging their audience during the early days of the week, suggesting the temporal interval to be mostly settled during working hours, preferentially before lunchtime. The commanded strategy also indicates how the weekends are avoided to perform the send of campaigns, due to the higher probability of reflecting lower engagement and undefined behavior patterns.

Understanding the distribution and usage patterns of the textual and visual modalities is essential to develop an in-depth characterization of the campaign, contributing as well to the analysis of user engagement. By examining the frequency of images and text tokens that compose the e-mails, valuable shreds of knowledge can be concatenated and employed to align the effectiveness of the multimodal affective computing task. Providing a solid foundation for understanding the relationship of the modalities to consumer behavior, Figures 4.5 and 4.6 provide a graphical perception of the modality's presence.

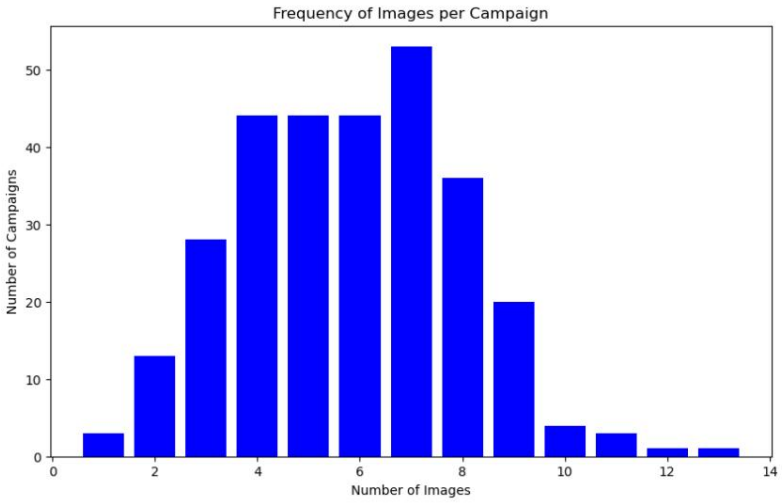

Figure 4.5: Frequency of Images per Campaign

Figure 4.5 reveals an important contribution to the analysis of the patterns regarding the number of images per campaign. The histogram showcases the frequency of employed visual segments, where most of the campaign uses between 4 to 8 images and the distribution relates to complementary information on the marketing routines. Most of the campaigns instantiate 6 images per campaign, being the highest frequency with more than 50 examples, suggesting a marketing tendency to moderately balance visual content in the developed campaigns. There are fewer campaigns with both sides of the established range, indicating that using too few or too many images may overload visual information and generate the opposite effect. Furthermore, on the image analysis, the content diversity was vast and presented some image samples where it was possible to encounter portions containing textual evidence. Given the disadvantage of not being able to align textual and visual modalities with one another, the objects and entities currently in the images held an in-depth analysis.

These included a large number of images with textual segments, with different sizes, colors, and formats, which give rise to potential techniques for image preprocessing.

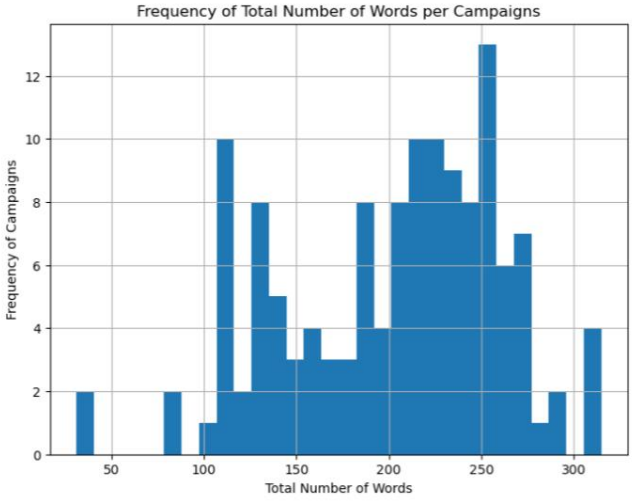

Figure 4.6: Frequency of Text Tokens per Campaign

Figure 4.6 analyzes the textual exhibits frequency per campaign. This distribution is granted by counting the number of words of the whole campaign text segments, in the x-axis and checking the frequency of campaigns, in the y-axis. The results of the campaign analysis show significant protrusions between 100 and 250 words per campaign, implying a reasonable but restrained use of text to enhance the interaction with the client. At the same, the minimalist presence of the campaign exhibiting too many or too few utterances indicates the effort of marketeers to balance the modality used, without creating overwhelming or unpleasant experiences with the target audience.

In conclusion, the exploratory data analysis of the accessible dataset has yielded important insights regarding marketing campaigns and their content by revealing the significance of the tabular, image, and text variables. The visual examples indicated that a reasonable amount of graphics was preferred to draw users' attention without being overpowering, while the text evidence suggested effective yet succinct content, usually in the range of 100 to 250 words. Deeper insights into campaign performance were also provided by the examination of openings, clicks, and unsubscriptions, which exposed patterns in user involvement and disengagement. Comprehending the impact of timing on interaction across various metrics also required an examination of temporal data, featuring the time of campaign sends. These results highlight the importance of having material that is balanced across modalities and paying close attention to temporal considerations, enhancing the forward steps of the experimental design.

4.2.2 Data Preprocessing

The data preprocessing stage for the E-mail Campaign Dataset is a critical step in preparing the raw data for effective analysis and modeling to fulfill the main objectives. This process involves cleaning, transforming, and structuring the data to ensure consistency and quality across all variables and turn them into modalities — text, images, temporal information, and key report metrics like sends, opens, clicks, and unsubscriptions. The goal of this stage is to handle missing values, solve inconsistencies, and format the data into a usable structure that aligns with the objectives of multimodal sentiment analysis.

According to the present research experimental design, a set of cleaning techniques is primarily administered to handle outliers and data discrepancies among the available sample of campaigns. As previously mentioned, the procedures begin by employing predominant variables that determine if a campaign is valid or considerable for the expression of inherent emotional information. To achieve this, the cleaning process initiated with an overview of the report metrics and possible alternatives that could raise the target variable to an affective computing task. The inherent marketing strategy behind these metrics serves as numeric evidence of customer preference, inducing positive feedback on the overall receivers over the total number of clicks attending the number of opens. Commonly used as a key performance indicator (KPI), the Click-Through-Rate (CTR) is obtained by measuring the percentage of the division between the number of clicks and the number of opens, performing the evaluation of an advertisement's success. However, it is also possible to perceive other metrics that translate the opposite effect of the clicks, considering the unsubscription count as an opposite effect to maximum engagement.

$$\mathrm{CTR}_X = \frac{(clicks_X - unsubscriptions_X)}{opens_X} * 100 \tag{4.1}$$

To take advantage of the reporting of customer interactive behaviors with campaigns, the CTR variable is calculated to approximate the campaign captures and audience retention with emotional analysis in marketing domains. The percentual measure of successful user interactions, described in Equation 4.1, is obtained by subtracting the unsubscriptions clicks from the overall clicks, and relating the result with the total number of opens. Assisting in transforming the positive experiences with different ranges, the numerical representation promotes the continuity of exploiting potential emotional outcomes. Through the consolidated rate of clicks, the preprocessing can pursue more improvement layers, using it to locate possible outliers.

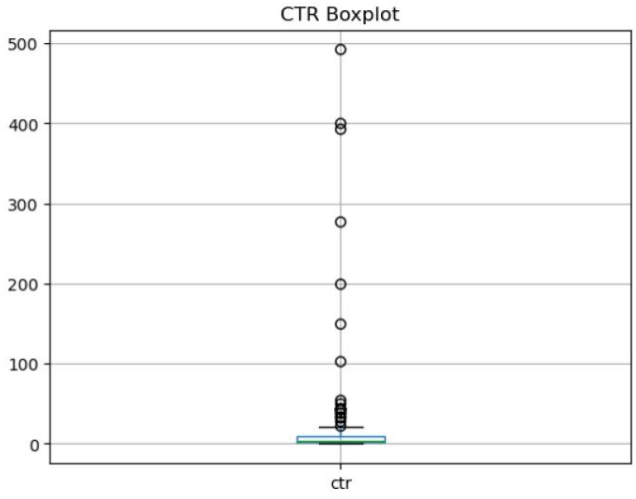

Figure 4.7: CTR Boxplot before Preprocessing

Figure 4.2 displays the CTR boxplot, which indicates that the majority of campaigns have poor click-through rates. The limited interquartile range and median show that most values are densely grouped at zero. However, a few campaigns surpass the clicks over the opens, indicating a stark disparity. Several outliers with disparate CTR measurements are provided. These campaigns influenced the remaining extracted variables as preprocessing went on and

were translated into testing campaigns and additional experimental submissions. Two main limitations are conveyed by the procedures for handling these discrepancies: while deleting these outliers considers a reduction in data size, substituting into average range values raises additional issues related to data augmentation techniques and the creation of synthetic data points.

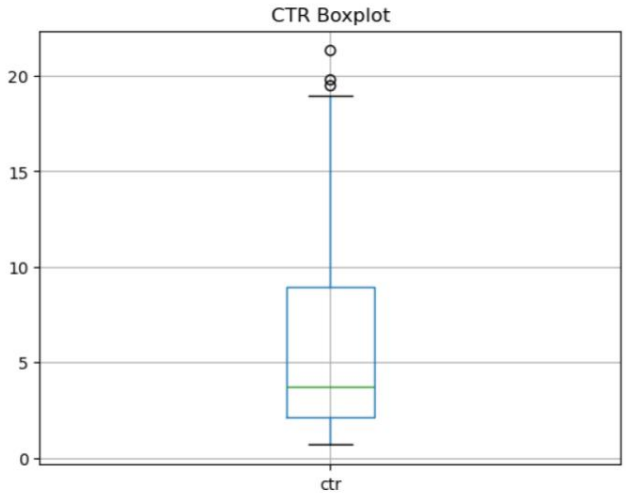

Figure 4.8: Boxplot of the Target Variable CTR

To handle these concerns, as well as prevent the model performance skewing, outlier removal techniques were applied, demonstrated in Figure 4.8. The complexity inherent in these data points implied a customized method to delete only the irrelevant campaigns, taking both manual analysis of each campaign and the appliance of winsorization. The statistical transformation while winsorizing extreme values is flawed by the fact that it does not take into account the reality of the business in campaigns where the CTR is admitted as valid, even if these campaigns are unique mass testing scenarios. The combined technique employed stabilizes the CTR distribution and improves the model's generalization by reducing the impact of unusual engagement.

After validating all the constraints with CTR variable, the admitted campaigns proceed to the preprocessing of text data. Term Frequency-Inverse Document Frequency (TF-IDF) and Natural Language Toolkit (NLTK) were used in several stages of the preprocessing of text variables to modify and clean the data in preparation for analysis. The frequency and structure of the analyzed text data are depicted in Figures 4.9 and 4.10.

Based on the frequency of text tokens per campaign presented in Figure 4.9, most campaigns have between 40 and 60 tokens, much less than the previous text analysis. This focus implies that most marketing communications are brief and direct, perhaps meant to grab attention fast. Campaigns with larger token counts, which correspond to lengthier and more in-depth content, are also shown by the distribution. Terms like typical marketing slogans that were more often used but less important across all campaigns were down-weighted after TF-IDF transformation, such that only contextually meaningful words made it into the final model. By using this method, campaigns' unique or highly relevant terms may be found, increasing the text-based sentiment analysis's accuracy.

Prospecting the assembling of relevant tokens, Figure 4.10 presents the word cloud of the tokens' frequency. The illustrative chart highlights the prevalence of terms like "peça,"

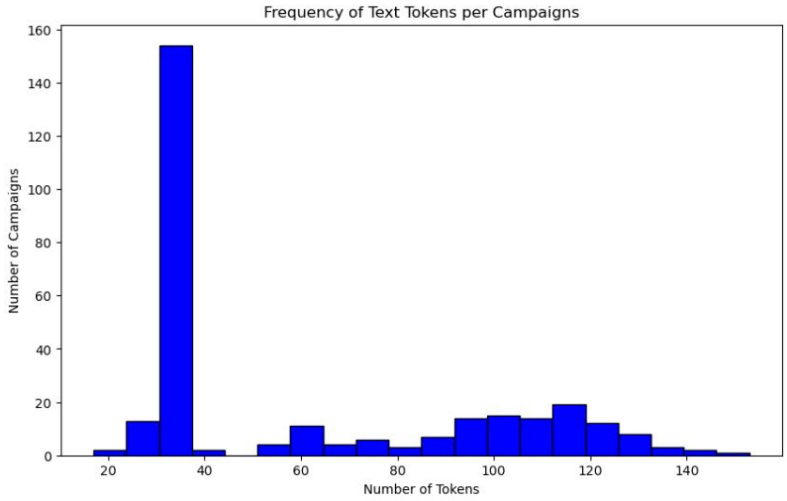

Figure 4.9: Frequency of Popular Text Tokens per Campaign

Figure 4.10: World Cloud of Text Tokens by Frequency

"você," and "look" by visualizing the most commonly used tokens across all campaigns. These terms draw attention to how frequently specific instructions and tailored marketing language are used. Tokenization, stop word elimination, and lemmatization using NLTK were among the preprocessing procedures that made sure only significant words were included in the study. The outcome is a more concentrated and organized word cloud that emphasizes the main marketing message, devoid of outside tokens and terms that don't add value to emotional research or campaign assessment.

To ensure improved performance in the upcoming sentiment and multimodal analysis stages, the text data has been efficiently cleaned, tokenized, and transformed by implementing these pretreatment approaches. Combining TF-IDF for significance weighting with NLTK for basic text cleaning allowed for the removal of noise and improved dataset optimization for enhanced insight creation. Yet, the preprocessing of textual utterances needed more understanding of the usage impact of each token to elevate the click rate. As so, Figures 4.11 and 4.12 display other text tokens preprocessing method by evaluating the presence of each token in campaigns with the highest CTR value, checking differential evidence between

frequency and popularity.

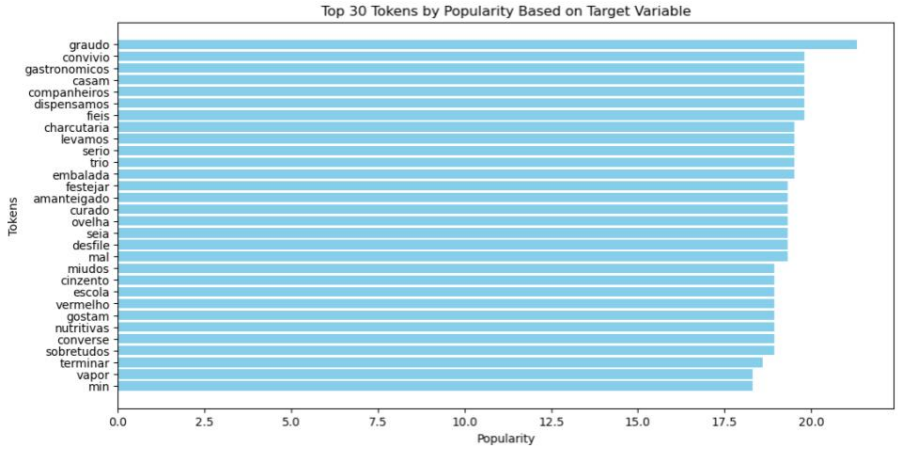

Figure 4.11: Frequency of Popular Text Tokens per Campaign

The horizontal bar chart, represented in Figure 4.11, further supports the word cloud by displaying the top 30 tokens ranked by their popularity based on CTR. Tokens like "graúdo," "convívio," and "companheiros" once again appear at the top, reaffirming their influence in high-engagement campaigns. The horizontal bars quantify the popularity of each token, providing a clear indication of which words are most strongly associated with successful campaigns. This visual reinforces the observation that certain terms, likely reflecting a specific tone or theme, consistently appear in campaigns that drive better performance.

Figure 4.12: World Cloud of Text Tokens by Popularity of CTR

Figure 4.12 displays the most popular tokens in campaigns with strong click-through rates. Given that the audience is likely to connect with these terms, marketing communications including them may have done a better job of promoting user involvement. Tokenization, lemmatization, and stopword removal were among the preprocessing techniques used in this case to make sure that only pertinent and significant tokens were taken into account for analysis. This makes it possible to comprehend the powerful language employed in high CTR advertising more clearly. This preprocessing pipeline, in contrast to general token

frequency analysis, focused on the relationship between text tokens and their effect on CTR performance. This allowed it to identify the words that drive user interactions in marketing campaigns and provide a weighted insight on the importance of each text token.

4.2.3 Feature Extraction and Modality Representation

The feature extraction procedure for the text and picture modalities, along with the goal variable, Click-Through Rate (CTR), are described in this section. The experimental setting relies heavily on the retrieved features from various modalities, which enable an integrated sentiment assessment approach with multimodality that can forecast campaign success in terms of CTR performance.

For the modality representation of textual observations, text embeddings were produced using the BERTimbau [66] model, a Portuguese-based BERT language model. The choice of this model, incorporating language semantical analysis, approaches the combination of the subject, the snippet, and the utterances found during the HTML content extraction. While using NLTK approaches to eliminate extraneous characters, punctuation, and stopwords, the BERTimbau tokenizer is capable of recognizing these elements and easily considering language barriers as relevant insights. By capturing semantic subtleties, these embeddings enable the model to identify minor patterns in the text, with a focus on emotive cues and persuasive language that are meant to increase user engagement.

The VGGNet was used to create image embeddings for the visual modality. After being processed to ensure a constant input size, the campaign-extracted images were run through the chosen model, which had previously been trained on ImageNet. To guarantee consistency in the feature extraction procedure, each image was normalized and scaled before it was uploaded into the model. The output obtained from VGGNet's second-to-last layer was used to create the visual embeddings. Critical characteristics including colors, forms, and objects are captured by these embeddings, which can influence CTR performance and increase user engagement. Visual components are essential for drawing attention in marketing efforts, and VGGNet-16 helps us recognize how important these components are.

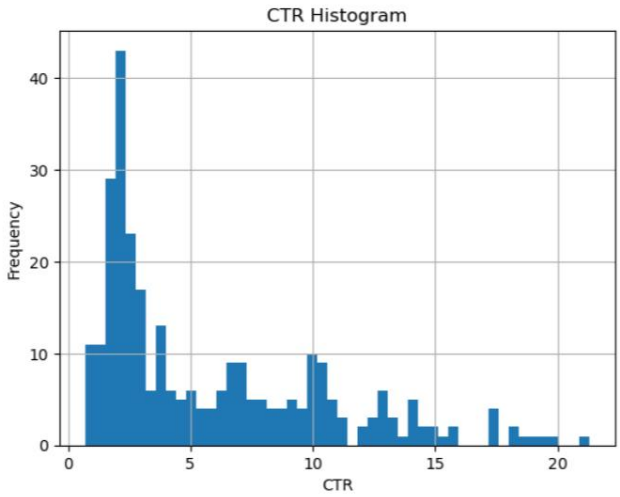

Figure 4.13: Histogram of the Target Variable CTR as Numeric

After preprocessing, the Click-Through Rate (CTR) histogram displays a more balanced and concentrated distribution, with most campaigns lying between one and three percent of the

total CTR. A more uniform portrayal of campaign success has been achieved by removing the extended tail of high CTR numbers above 20% by the elimination of outliers. Though they are fewer in number, higher-performing advertisements with CTRs between 5 and 10% are nevertheless occasionally seen. In the end, more accurate analysis and predictions result from this modified distribution, which guarantees more dependable input for modeling and reflects overall user interaction patterns without being distorted by exceptional outliers.

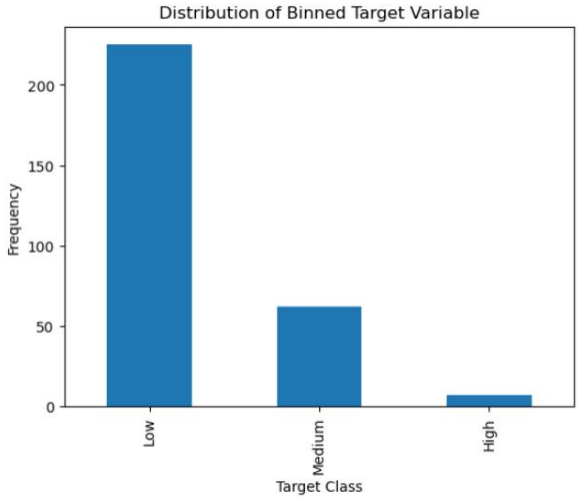

Figure 4.14: Frequency of the Target Variable CTR as Multi Class

The distribution of the CTR target variable among the three classes is depicted in Figure 4.14. With very few in the "High" class and most falling into the "Low" or "Medium" classes, it can be concluded that most campaigns had comparatively poor click-through rates. The classification process may be impacted by this distributional imbalance, which necessitates the use of strategies like sampling or weighting methodologies to solve the unequal class distribution.

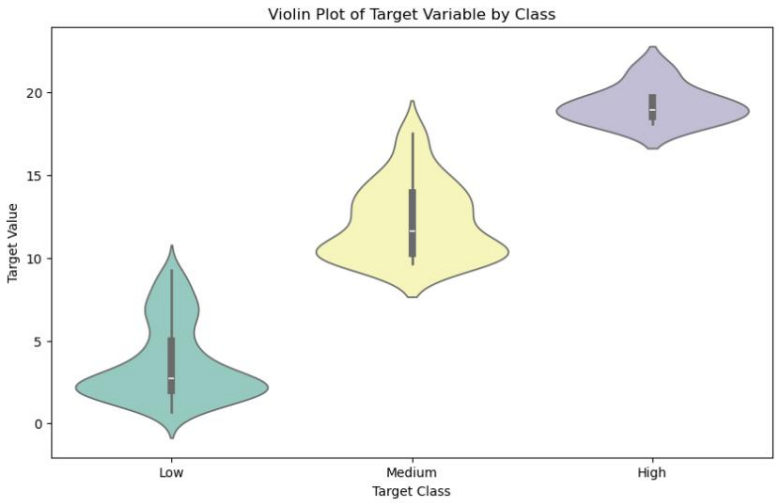

Figure 4.15: Violin Plot of the Target Variable CTR as Multi Class

The target value distribution within each class is shown in detail through the violin plot in Figure 4.15. The discrete shapes of the three classes - Low, Medium, and High - indicate the distribution of CTR values over the range. The interquartile range (IQR), with a distinct

division between the classes, is displayed in the middle box plot inside each violin. This ensures that the binned target values are appropriately distinguished for the classification job by confirming that the target variable has been evenly distributed among the bins with no overlap.

Table 4.2: E-mail Campaign Dataset Statistics

Statistics	**Values**
Total number of samples	294
Total number of images	1 710
Total number of sentences	1 743
Total number of words in sentences	61 925
Total number of relevant tokens	18 002
Total of unique words in sentences	2 311
Average number of images in a campaign	5,8
Average number of sentences in a campaign	5,9
Average word count per sentence	210,6
Total number of words appearing at least 10 times in the dataset	386
Total number of words appearing at least 20 times in the dataset	258
Total number of words appearing at least 50 times in the dataset	156

Understanding the size and complexity of the data utilized in the study requires a comprehension of the composition and features of ECD, which is fully summarized statistically in Table 4.2. There are 294 examples in the collection, with 1,710 pictures and 1,743 phrases altogether. The dataset may be subjected to multimodal analysis, which combines textual and visual data, as shown by the comparatively high quantity of phrases and images.

The dataset contains 61,925 words with textual content; 18,002 of those words are recognized as meaningful tokens, indicating the large amount of textual material that may be processed. Additionally, the dataset exhibits lexical variety, with 2,311 distinct terms found in the phrases. A balanced integration of written and visual content is evident across samples, with an average campaign consisting of 5,8 images and 5,9 phrases. With an average word count of 210,6, the sentences appear to be quite lengthy, most likely as a result of intricate details or complicated subject matter. Additionally, the table offers insightful information on word frequency, which is important for a variety of textual processing applications. In particular, 386 terms appear at least ten times, 258 words that appear at least twenty times, and 156 words that appear at least fifty times. These frequency counts are crucial for training models, extracting features from the data, and making sure the data is rich enough in words that appear frequently enough to enable reliable machine learning models.

Overall, the dataset's diversity in terms of images, sentences, words, and lexical richness highlights its suitability for in-depth analysis, particularly in applications requiring multimodal data fusion and natural language processing. This statistical overview concludes the possibility of modulating the collection into an affective computing task by emphasizing the dataset's relevance and potential for producing real, data-driven outcomes.

4.3 Use Case 2: MOSEI Dataset

The process of evaluating, preparing, and converting the MOSEI dataset into a reliable multimodal sentiment prediction resource is described in this section. The dataset aims to capture the subtle emotional emotions in various video segments and includes textual, visual, and audio modalities.

4.3.1 Exploratory Data Analysis

The MOSEI dataset, one of the largest multimodal sentiment analysis datasets, contains rich information across three main modalities: text, visual (video), and acoustic (audio). This dataset features over 23,000 video segments annotated for sentiment and emotion, enabling a thorough investigation of the relationship between multiple modalities and sentiment expression. An initial exploratory analysis of the dataset provides critical insights into the distribution of the data, as well as the interaction between modalities.

For the textual modality, transcriptions of spoken words are analyzed by examining word frequency distributions and understanding sentiment-rich tokens. A word cloud visualization highlights common tokens used in the dataset, allowing us to identify specific words that frequently occur in positive, negative, or neutral sentiment expressions. Token frequency histograms further clarify the distribution of text length across samples, revealing that most segments contain a moderate number of tokens, which helps inform decisions about padding and truncation for text processing.

For the visual modality, an analysis of facial expressions and actions captured from video frames allows us to understand the role of visual cues in sentiment expression. Histograms of frame lengths and summary statistics for facial Action Units (AUs) provide insight into the visual richness of each video segment. Additionally, initial checks of the audio modality involve looking at the distribution of vocal intensity, pitch, and other acoustic features, which further adds to our understanding of emotional cues present in speech.

The dataset's sentiment labels, provided on a continuous scale from -3 to +3, are binned into negative, neutral, and positive categories for a better understanding of their distribution. A visual examination of sentiment and emotion class distribution helps inform us about potential class imbalance, with negative and neutral sentiments occurring more frequently. This provides early warnings for potential model bias and informs the need for balancing techniques.

4.3.2 Data Preprocessing

Data preprocessing is an essential step to transform the raw MOSEI dataset into a format suitable for Multimodal ML tasks. Each modality requires specific preprocessing strategies to extract meaningful features and ensure consistency across modalities.

For the text modality, preprocessing includes tokenization, lowercasing, and stopword removal to clean the transcripts and reduce noise in the data. Lemmatization is applied to reduce words to their base form, allowing similar words to be grouped together. This step ensures that each word's meaning is represented in the context of the surrounding text, a crucial factor for sentiment analysis.

The preprocessing for the visual modality involves frame extraction from video segments. The videos are split into individual frames, with identifier keys for each frame and regular

intervals selection. Each frame undergoes several procedures, including face detection and alignment to standardize the orientation and position of faces. This is a critical benefit for ensuring consistent facial feature extraction to accurately predict the affective label.

For the acoustic modality, the preparation of data begins with Voice Activity Detection (VAD) to identify and segment speech regions in the audio. This ensures that irrelevant sections, such as silence or background noise, are excluded from the analysis. Acoustic features, such as MFCC, pitch, and intensity, are extracted to capture the emotional content of the speech. These features are normalized to account for varying speaker volumes and acoustic environments across the dataset.

Finally, the general data synchronization across modalities is performed by aligning the textual, visual, and acoustic data based on the video timestamps. This ensures that the features from different modalities correspond to the same time segments, enabling coherent multimodal analysis.

4.3.3 Feature Extraction and Modality Representation

Feature extraction and embedding generation from the three modalities enables the development of a multimodal model that leverages information from different sources to enhance sentiment analysis. The availability of this dataset is ensured by the Software Development Kit (SDK) developed by the MultiComp Lab to simplify the use of data collections and the pre-trained models, being a powerful tool to study multimodality and affective computing. However, the included modalities are already preprocessed, available for any researcher in its embeddings format, due to ethical concerns of the included speakers. This particularity grants to the users the possibility to add computational sequences to improve the label prediction, aligning the required modality to the specific label, focusing on the customized use of this dataset.

The textual embeddings are generated from the pre-trained BERT model and serve as the primary representation. BERT embeddings are highly context-sensitive, capturing not only the meaning of individual words but also their relationships with surrounding words. This allows for nuanced sentiment predictions, where subtle changes in wording or sentence structure can alter sentiment interpretation. The embeddings provide a dense, continuous representation of each token, contributing to the final sentiment prediction.

Pre-trained Openface is used to extract high-level features from video frames, designing the visual embeddings, representing facial features, capturing expressions such as smiling, frowning, and other cues that are essential for detecting emotion. Since visual data is inherently temporal, sequences of these embeddings across frames are utilized to capture changes in expressions over time, providing a richer understanding of sentiment dynamics. Facial action units (AUs) are also tracked, providing explicit measures of facial muscle movements that correspond to emotional states.

For the audio modality, extracted features like MFCCs, pitch, and intensity form the basis of acoustic embeddings generated by using COVAREP. These features capture the emotional undertones of speech, such as excitement, anger, or sadness, which are often reflected in changes in voice pitch or loudness. By combining acoustic embeddings with textual and visual representations, the model can consider not only what is being said but also how it is said and expressed.

By integrating text, visual, and acoustic features, along with the multi-class distribution of sentiment labels, the feature extraction process ensures that the final model has a comprehensive, multimodal representation of the data. This enables a more holistic approach to predicting sentiment in user engagements present in the MOSEI dataset.

Table 4.3: MOSEI Dataset Statistics based on the MultiComp Lab official website [67]

Statistics	**Values**
Total number of sentences	23 453
Total number of opinion sentences	18 148
Total number of objective sentences	5305
Total number of videos	3 228
Total number of distinct speakers	1 000
Total number of distinct topics	250
Average number of sentences in a video	7.3
Average length of sentences	7.28 seconds
Average word count per sentence	19.2
Total number of words in sentences	447 143
Total of unique words in sentences	23 026
Total number of words appearing at least 10 times in the dataset	3 413
Total number of words appearing at least 20 times in the dataset	1 971
Total number of words appearing at least 50 times in the dataset	888

An extensive summary of the statistics from the CMU-MOSEI dataset is provided in Table 4.3, which was retrieved from the MultiComp Lab website. This multimodal dataset, which consists of textual, visual, and audio inputs, is extensively utilized for sentiment analysis and emotion detection applications. The dataset's wide diversity in both subjective and objective content is seen in its 23,453 phrases, of which 18,148 are opinion-based and 5,305 are objective.

As evidence of the wide range of subjects covered by the dataset, 3,228 films overall that record interactions from 1,000 different speakers covering 250 different themes are included in the collection. For tasks like multimodal sentiment analysis or speaker recognition that need a speaker and topic diversity, this variation is essential. Videos have an average of 7.3 sentences per video, with an average sentence length of 7.28 seconds, highlighting the dataset's extensive temporal and content coverage. The dataset has 447,143 words in total, with an average sentence length of 19.2 words, indicating a moderate sentence length. With 23,026 unique terms in the dataset, its lexical richness and potential for creating reliable natural language processing models are highlighted. Word frequency statistics indicate that the dataset is well-suited for machine learning models that need repeating keywords for training and feature extraction, with 3,413 words appearing at least 10 times, 1,971 words appearing at least 20 times, and 888 words appearing at least 50 times.

In summary, the MOSEI dataset, with its varied variety of speakers, themes, and well-distributed sentence structure, is a significant resource for researchers conducting thorough multimodal analyses. This dataset is crucial for raising the bar in human-centered AI research as well as the Affective Computing integration since it provides a rich multimodal basis, which

emphasizes the dataset's capacity to provide scalable, precise, and insightful models across a range of analytical disciplines.

4.4 Experimental Setup

The primary focus in the Experimental Setup is the development and execution of every detail considered on the multimodal affective computing framework, combining various ML and DL techniques. As outlined in Section 3.4 where the Experimental Design is described, the setup is assembled accepting different datasets, each one with a distinct expertise level and contribution, to generalize the results on the framework and ultimately be implemented as a real-time tool for different marketing contexts.

The first set of the experimental setup involves the preprocessing of each modality, considering the overall customizations. Textual data from ECD is obtained by joining all the text variables, such as the subject, the snippet, and the HTML text, and preprocessed using TF-IDF, to analyze the frequency of each word and create the average value of text popularity. This metric is calculated to enhance the alignment between textual features and the target variable. NLTK is also employed, having the real version and a Portuguese-based one to remove stop-words and tokenized textual elements, handling the inherent language issues before using them as model input. For improved tokenization and embedding generation, the BERT model was also highlighted due to the robustness of embedding generation and classification tasks. In this case, the extraction of special tokens like CLS and SEP was discarded from the ECD textual modalities due to the embeddings being generated word-level and not sentence-level. The tabular data was also preprocessed, using both weekdays and time enconding to include those features in the experimental model performance. The MOSEI dataset considers the sentence-level embedding generation, creating strategies enhanced in the diversity of approaches to evaluate during the model performance comparative analysis. On the other hand, image data is processed using VGGNet to extract image embeddings, and tabular data is encoded using temporal encoding techniques. For the MOSEI dataset, audio features are extracted using COVAREP, which is designed to capture acoustic properties relevant to emotional expression. All preprocessing steps ensure that the data is structured and ready for the next phase of embedding generation and fusion.

The experimental setup adopts two main strategies for integrating multiple modalities: Concatenation for Implicit Feedback on the ECD dataset and Late Fusion for Explicit Feedback on the MOSEI dataset. In the case of implicit feedback, the tabular, textual, and visual embeddings from the e-mail campaigns are concatenated in the late stage of the framework. This combined feature set is then fed into a series of machine and deep learning models, including Decision Tree (DT), K-nearest neighbors (K-NN), Support Vector Classifier (SVC), RF, Gradient Boosting (GB), and developed Multimodality Neural Network, stated as MultiNN. These models are trained to predict the implicit emotional response, using the CTR (Click-Through Rate) as the baseline for the categorical multi class target variable. For the MOSEI Dataset, a late fusion strategy is employed, where the text, image, and audio embeddings are first processed independently and then combined in a late fusion stage. The combined multimodal embeddings are used to predict the explicit emotional feedback based on the sentiment valence labels provided in the dataset.

Once the models are trained, their performance is evaluated based on metrics such as Precision, Recall, and F1-Score. The results for each model are compared across different datasets, with a focus on understanding how the fusion of modalities affects predictive

accuracy. The experimental results are particularly insightful for understanding how well the proposed setup performs in both implicit opinion and emotional state due to the customer behavior, and explicit feedback opinion and emotional state due to the customer expressions. After initial evaluations, hyperparameter tuning and model optimization are performed to improve the accuracy and efficiency of the best-performing models. Techniques like cross-validation and grid search are applied to select optimal parameters for each model, ensuring robust and generalizable results across both datasets.

The experimental setup focuses on effectively managing and fusing the multimodality from both datasets to evaluate the model performance, examining how different modalities contribute to the prediction of emotional responses. The results and performance of the models are further explored in the following sections, leading to insights into the benefits and challenges of multimodal sentiment analysis in marketing and affective computing tasks.

4.5 Summary

This chapter focused on detailing the procedures of data analysis, preparation, and modeling for the ECD and MOSEI datasets, highlighting their multimodal structures. The preprocessing techniques involved cleaning and organizing data, extracting features, and generating embeddings using advanced models like TF-IDF, NLTK, and BERT for textual evidence and VGGNet and OpenFace for images, while MOSEI prepared audio features using COVAREP. Additionally, the experimental setup was outlined, concentrating on the integration of multiple modalities through both implicit and explicit emotional feedback models. The setup involved concatenating embeddings from different modalities and employing Machine and Deep Learning methods, optimizing the prediction of the target variable for affective computing. This approach ensured a comprehensive evaluation of model performance, leveraging both datasets to maximize classification accuracy and predictive power across different emotional contexts.

Chapter 5

Results and Discussion

The outcomes of the applied methods and the evaluation of the results from the experiments carried out on the ECD dataset and the MOSEI dataset are discussed throughout the Results and Discussion chapter. In addition to describing the prediction models' performance across multiple modalities, this chapter offers a thorough explanation of the findings within the conceptual structure of multimodal sentiment analysis. This section offers explanations for the success of the sentiment prediction task, identifies the difficulties faced, and investigates the ramifications of the results in practical marketing scenarios by contrasting the efficacy of the suggested strategy.

5.1 Experimental Results and Evaluation

The results of this study present the performance of the proposed experimental setup approach on the ECD and MOSEI dataset. The model appliance procedure evaluated the modality integrity and the multimodal representation of the target feature on both collections. Due to disparities in the structure and content of each dataset, the results of the prepared study environment consider several concerns regarding the model performance for the two datasets. The results on the MOSEI dataset are a corresponding baseline for the similar characteristics on the E-mail Campaign dataset, due to the transformation enforcement on the last to some routines of the first. Therefore, the experimental results explore the model performance to validate one of the best alternatives for sentiment valence prediction, comparatively to the results obtained with the MOSEI.

Using precision, recall, and F1-Score metrics as benchmarks, we provide the experimental outcomes of applying several machine learning models to the ECD and the MOSEI Dataset. As known, the target variable distributes between 3 classes - low, medium, or high - having the ECDe dataset variant, where it is evenly distributed over the labels, and ECDr, where it is range-distributed to predict continuous engagement labels, are the two tasks that make up the ECD. Textual and visual modalities (T+V) are used in both tasks, while MOSEI employs textual, visual and audio modalities (T+V+A), as shown in Table 5.1

For the ECDe task, where the target variable was equally distributed across three engagement levels (low, medium, and high), the Multi-Layer Neural Network (MultiNN) model emerged as the top performer with a precision of 80.13%, recall of 78.96%, and an F1-Score of 72.85%. The Gradient Boosting (GB) model also demonstrated strong performance with a precision of 79.67%, recall of 75.82%, and an F1-Score of 69.46%, while the Random Forest (RF) model showed higher precision at 76.30% but a lower F1-Score of 56.83% due to its lower recall. Support Vector Classifier (SVC) achieved a balanced performance with an F1-Score of 70.81%, precision of 72.55%, and recall of 69.67%. Traditional models

Table 5.1: Experimental Results on E-mail Campaign and MOSEI datasets

Dataset	Model	Precision (%)	Recall (%)	F1-Score (%)
ECDe (T+V)	DT	66,53	68,54	67,50
	K-NN	69,10	70,78	66,52
	SVC	72,55	69,67	70,81
	RF	76,30	68,54	56,83
	GB	79,67	75,82	69,46
	MultiNN	80,13	78,96	72,85
ECDr (T+V)	DT	54,77	40,44	42,22
	K-NN	67,05	64,04	64,85
	SVC	68,32	60,67	62,38
	RF	64,42	59,55	61,16
	GB	63,03	58,43	59,86
	MultiNN	68,05	61,84	63,59
MOSEI (T+V+A)	DT	71,82	75,52	74,49
	K-NN	78,35	75,15	76,54
	SVC	79,21	76,87	77,51
	RF	82,33	79,56	80,64
	GB	83,01	78,88	80,86
	MultiNN	85,01	82,75	83,69

like K-NN and Decision Tree (DT) lagged behind, with F1-Scores of 66.52% and 67.50%, respectively. These results indicate that more complex models such as MultiNN and GB are better suited for handling the interaction between text and visual modalities in equally distributed sentiment prediction tasks within marketing communications.

In comparison to ECDe, the model's performance was often poorer for the ECDr task, where the target variable had a ranging distribution and was used to predict engagement metrics like Click Through Rate (CTR). This was likely due to the more challenging regression-based predictions. With an accuracy of 68.05%, recall of 62.84%, and an F1-Score of 63.59%, the MultiNN model maintained its dominance. Models like GB and RF, on the other hand, performed worse; GB achieved 63.03% precision, 58.43% recall, and a 59.86% F1-Score, while RF achieved 61.16%. With an F1-Score of 62.38%, the SVC model did somewhat better than in ECDe, although it was still below MultiNN. Less complicated models such as K-NN and DT have trouble with the task's intricacy, achieving F1-Scores of 64,85% and 42,22%, respectively. Overall, the results suggest that ranged-distribution is more challenging, requiring more advanced model architectures to achieve satisfactory results with unbalanced data.

The models performed best on all tasks in the MOSEI dataset, which combines text, visual, and audio modalities for sentiment analysis. With an accuracy of 85.01%, recall of 82.75%, and an F1-Score of 83.69%, the MultiNN model surpassed the competition once more, demonstrating its effectiveness in handling complicated multimodal data. Both RF and GB demonstrated strong performance; GB achieved an F1-Score of 80.86%, recall of 78.88%, and precision of 83.01%. The SVC model came in second with a solid 77.51% F1-Score. Achieving an F1-Score of 76.54%, K-NN outperformed simpler models like DT, which had

the lowest F1-Score of 74.49%. The inclusion of audio data in MOSEI, alongside text and visual modalities, significantly enhanced the performance of the models, demonstrating the importance of multimodal fusion in improving sentiment and emotion recognition tasks.

5.2 Comparative Evaluation and Discussion

The experimental results from the ECD and the MOSEI dataset demonstrate the significant influence of the data preprocessing, transformation, modality representation, and embedding generation processes on model performance. These steps, which are critical in preparing the data for input into machine learning models, directly impacted how well the models were able to handle sentiment and engagement prediction tasks across different modalities.

The experimental setup, which used both ML and DL models for the learning approaches, underscored the importance of multimodal fusion for sentiment prediction. In the ECD, the best-performing model was MultiNN and achieved the highest results, but the overall performance was constrained by the limited emotional depth of the text and visual data. ML models like RF and GB performed relatively well but struggled with the ranged distribution of the target variabke in ECDr, where the prediction of continuous engagement metrics was more difficult due to the limited emotional information in the dataset.

Overall, the comparative analysis of the results reveals that the quality of data preprocessing, transformation, and modality representation has a direct impact on model performance. In the ECD, the lack of emotional depth and modality diversity limited the ability of the models to accurately predict sentiment and engagement, particularly in the ECDr task, where the target variable was distributed according to the CTR range, lacking on balancing data. In contrast, the MOSEI Dataset, with its rich emotional annotations and multimodal inputs, allowed the models to perform significantly better, particularly in sentiment classification tasks. The MultiNN model consistently outperformed other models, particularly when working with multimodal data that included text, visual, and audio inputs.

5.3 Domains Implications

The results obtained from the ECD and the MOSEI datasets have significant implications for the use of multimodality and affective computing fields in the marketing domains. The comparative discussion highlights the importance of powerful, multimodal data and advanced model architectures in achieving accurate sentiment and engagement predictions, which can be directly applied to real-world marketing strategies and customer experience management.

In the marketing domain, the findings from the ECD suggest that while traditional text and visual-based analytics can offer insights into customer engagement, they are limited in their ability to capture deeper emotional signals. The reliance on basic engagement metrics like CTR and the use of static images provide a shallow understanding of customer sentiment. Models such as MultiNN performed well in this context but were constrained by the lack of emotional richness in the dataset. This underscores the need for marketing applications to incorporate more dynamic and interactive modalities, such as real-time audio or video feedback, to enhance the depth of customer insights.

Additionally, the results from the MOSEI dataset highlight the growing importance of affective computing in creating emotionally aware systems. The superior performance of the MultiNN model in handling multimodal inputs suggests that future applications of sentiment

analysis, particularly in domains such as virtual assistants, telehealth, and education, will need to integrate audio, visual, and textual data to provide more accurate and contextually relevant emotional insights. The strong results also support the idea that Affective Compuiting combined with AI can improve machine understanding of human behaviors, emotions, and intentions, making systems more empathetic and human-centric.

The comparative evaluation between the two datasets illustrates that multimodal fusion and the inclusion of rich emotional annotations are essential for capturing the complexity of human emotions in a variety of domains. In marketing, sentiment analysis based on just text and visual data is limited in depth, but when enhanced with audio and other dynamic modalities, the predictive power of models increases significantly. This suggests that for industries like marketing, entertainment, customer service, and healthcare, multimodal approaches should be prioritized to fully harness the potential of sentiment and emotion analysis.

The key implication is that businesses and systems that incorporate multimodality and affective computing will have a competitive edge by gaining a deeper, more nuanced understanding of their users or customers. This allows for the development of personalized, emotionally responsive systems that can adapt in real-time, providing more meaningful interactions and improving overall satisfaction and engagement. As technology advances, the ability to capture and process diverse forms of emotional input will become a cornerstone of successful customer experience management and human-computer interaction across multiple domains.

5.4 Summary

Featuring a spotlight on multimodal affective computing prediction, this chapter included the experimental findings and assessments of machine learning models applied to the MOSEI and the ECD datasets. Precision, recall, and F1-Score measures were used to evaluate the models' performance, demonstrating the significance of multimodal fusion, data preprocessing, and modality representation in enhancing prediction accuracy.

The MultiNN model outperformed other models in the ECD, which includes both textual and visual modalities. However, the dataset's lack of emotional richness hindered the model's overall performance, especially for the ECD range-distributed task that used range-distributed engagement measures. More sophisticated models, like Random Forest and Gradient Boosting, did well too, although they had trouble keeping up with the intricate nature of the classification problem. By comparison, the models were able to attain better accuracy thanks to the MOSEI dataset, which included text, visual, and audio modalities. Multimodal inputs with rich emotional annotations greatly enhanced the model's performance, with MultiNN outperforming all other measures. The addition of audio data was very helpful in showcasing the effectiveness of multimodal fusion in sentiment analysis assignments.

Such results draw attention to the critical importance of multimodal integration and strong preprocessing to improve model performance. The results have applications in the fields of affective computing and marketing, where richer multimodal data—including audio—can provide a more in-depth understanding of opinion and engagement. The study emphasizes how incorporating dynamic and varied modalities might improve the user experience and increase the depth of emotional analysis in future applications.

Chapter 6

Limitations and Future Directions

This section outlines the limitations during the exploratory development and implementation of the multimodal sentiment analysis prototype for marketing domains. While discussing all the results and procedures, highlighting significant issues in integrating diverse data modalities for affective purposes, it is also possible to prospect potential alternatives of future operations guidelines to improve the following outcomes. Amplifying limitations from the quality and diversity of datasets to the performance of the conceived models requires a critical perspective in formulating new solutions considering business interests and technical robustness. As such, the addressed future directions enhance compliance with the advanced practical architectures and the ethical and regulatory standards in Marketing and AI.

Domain Complexity

One of the key limitations in applying multimodality and affective computing to improve marketing strategies lies in the complexity of the domains that are applied to the developed technologies. The intricacy nature of each multidisciplinary field can be pinpointed as a misleading complement in obtaining meaningful outcomes, succumbing to a narrow spectrum of conclusions and impracticable deployment as a real-time service. Even if the domain-specific nuances of one research field grant a greater contribution to the developed work, it can not be ensured that the information can and will be relevant to other domains. While Marketing-empowered analytics plays a significant role in customer behavioral pattern breakdown not only for the evaluation of advertising strategies' success but also for the multimodal configuration of modalities, affective computing may need deeper insights into the collected data and following analytic reports. At the same time, leveraging affective computing in modality diversity challenges the generalization of customer opinion, not granting a more robust marketing tool to perceive flexibility to minority cases and elevate market competitiveness.

To overcome the posed challenges by domain complexity in applying multimodality and affective computing to marketing, the development of more adaptive and versatile models that can navigate through the intricate nature of their combined usage must be prioritized. Promising solutions based on domain adaptation and transfer learning techniques are used to optimize model training details on marketing projects to generalize more effectively across the other domains with domain-specific datasets. Ensuring flexibility to handle multimodal insights to their utmost relevancy while preserving crucial emotional annotations contributes significantly to dealing with distinct marketing environments. This can also be enhanced by advancing the synergy between trustworthy affective computing to marketing analysis, generating comprehensive and consensual datasets that provide more vigorous insights into consumer preferences. By integrating these capabilities, future tools will become more

competent in handling personalized marketing strategies and offering faster, more responsive interventions while keeping the customer's ethical and regulatory rights conserved.

Dataset Configuration

The configuration of datasets represents another crucial limitation in this research, particularly focusing on the quality, completeness, design, and diversity of the used data for multimodal sentiment analysis in marketing. As the baseline of the overall proposed solution, the way of capturing, shaping, processing, and aligning these collections introduces several issues that affect the performance of the models and the generalization of obtained results. As a multimodal-based project, the inherent heterogeneity of data and their sources increases the complexity, often conveyed by inconsistent data structures, underrepresentation of modalities, different granularities for the same modality, and difficulties in synchronizing with other modalities. These issues integrate alongside the affective computing challenging points, lacking ground truth of emotional annotations between different modalities and leading to an imbalanced arrangement of data, limiting multimodal and affective learning richness.

As a fundamental element for the overall research, dataset configuration requires focus on several key areas, linking the improvement of data collection and standardization as the first one. This process is essential for securing dataset consistency, enriching information, and a solid architecture for the following procedures of comparing different modalities and marketing contexts. Standardized data formats contribute to the mitigation of inconsistencies and the integration of multimodal inputs, enhancing alignment and fusion techniques for cohesive data samples. At the same time, the improvement of remarking emotional cues across all modalities can be achieved by advanced automatic tools to contribute to an accurate affective learning procedure with different modalities. Also, the upgrade of modality-specifics with the annotation of emotional signals can reduce the under- and over-representation of modalities, allowing flexibility to the study environment. Exploring the dataset dynamical tailoring to different marketing contexts with the incorporation of these elements boosts the systems stability to every scenario.

Multimodal Integration

Another key challenge in multimodality is the integration of different modalities, presented in this research with the textual, visual, audio, and tabular data from the 2 datasets. Combining diverse formats of data like these into a unified multimodal representation and proceeding to its practical integration reveals several limitations that affect the framework's performance. Among them, the discrepancy between modalities underlined in the granularities and the data structure that compromise the alignment process. While each provides complementary information to the other, conceiving a suitable connective relationship among them is challenging, due to the lack of different levels of detail that create redundant cues and biased samples. Simultaneously, an incorrect multimodal integration originates inappropriate modality weighting, overly relying upon one modality based on its presence against the absence of the other, to an incomplete sentiment analysis. Also directly connected to inaccurate affective computing is the feature fusion approach, having particularities of performance with traditional and advanced techniques. Adjusting modalities according to the chosen methods can lead to irrelevant usage of methods prone to overfitting when not properly optimized.

Future studies should concentrate on sophisticated feature fusion methods, such as co-attention networks and cross-modal attention mechanisms, which may dynamically capture links across modalities and lessen bias or redundancy, to solve the limits of multimodal integration. Furthermore, improving weighting and modality adaption procedures will guarantee that the contributions of each modality are suitably distributed according to their importance, avoiding an over-reliance on any independent modality. Lastly, by enhancing temporal and semantic alignment techniques like temporal synchronization and semantic mapping, all modalities will be perfectly matched, resulting in sentiment analysis that is more coherent and precise. These strategies will improve the efficacy of emotional computing in marketing by decreasing overfitting, optimizing multimodal integration, and improving overall performance.

Technical Performance

The technical performance of personalized multimodal sentiment analysis models is a determining factor in evaluating the effectiveness and scalability of the proposed solution. The efforts to keep up with the advances in model architectures and computational resources are not always fulfilled, particularly due to real-world marketing applications. Combining multiple data types takes a high computational cost on the overall development, from training machine and deep learning models to fine-tuning, resulting in an unfeasible approach for real-time and large-scale applications. As previously inferred, data inputs can also affect model complexity, increasing the risk of overfitting when working with smaller, less specified training datasets and using multimodal transformers that require generalized context-based collections. The inability to capture the concept of different domains and the latency to process them hinders the model sensitivity to deploy scalable and in time returns to become an applicable assistance to multimodal sentiment analysis.

Future research should examine model compression and efficiency optimization techniques, such as pruning, quantization, and knowledge distillation, to reduce the size and complexity of models without compromising performance. This will help address the technical performance limitations in personalized multimodal sentiment analysis. These methods will reduce computing costs by enabling models to function well in large-scale and real-time marketing applications. Furthermore, by refining models already optimized for multimodal tasks, transfer learning and pre-trained models can speed up development by lowering the requirement for large amounts of training data and improving model generalization. Overfitting may be further avoided by putting regularization and optimization strategies like dropout, early halting, and L2 regularization into practice, especially when dealing with smaller datasets. Ultimately, by integrating domain adaptation and meta-learning techniques to improve cross-domain resilience, models will be able to adjust to different marketing scenarios, increasing flexibility and decreasing delays in the processing of domain-specific inputs for more timely and sensitive sentiment analysis.

Ethical and Regulatory Compliance

Ensuring data privacy and anonymization is a major constraint on the use of multimodal sentiment analysis for marketing, particularly when handling sensitive data like faces or voices in photos and audio. Textual data is readily anonymized, but multimodal inputs from social media and consumer contacts raise the possibility of breaking privacy laws. Furthermore, multimodal models trained on huge datasets may inherit biases related to gender, color, or socioeconomic status, which can distort marketing decisions and result in uneven treatment

of client groups. As a result, prejudice and fairness continue to be major concerns. Another level of complication arises from ensuring compliance with laws like the AI Act and the GDPR, which impose stringent rules on the implementation of AI systems and information processing. Moreover, multimodal models' lack of explainability and transparency raises ethical questions by making it challenging to defend the choices these models make, which impedes accountability and justice in marketing applications.

The next step should concentrate on designing sophisticated privacy safeguards such as differential privacy and strengthened anonymization methods to meet these constraints and guarantee that sensitive multimodal data is safeguarded without sacrificing its usefulness in sentiment analysis. To make sure that models handle all client groups equally, strategies for bias identification and mitigation need also be improved through approaches like fair representation learning and bias audits. Future work should also put a high priority on ensuring that real-time data processing complies with legal requirements, as well as incorporating frameworks such as the AI Act and GDPR into model development pipelines. In an attempt to raise trust and accountability in marketing practices, future research must explore the use of explainable AI (XAI) techniques like Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic explanations (LIME) to provide clearer insights into how multimodal models make decisions.

6.1 Summary

In summary, this chapter outlines the major constraints and potential guidelines in the application of multimodality and affective computing for marketing. The challenges identified include the complexity of handling domain concepts, configuring robust heterogeneity in the datasets by integrating different modalities, and maximizing the technical performance to its fullest. Ethical and regulatory concerns are another critical aspect of the developed work that, alongside the previous ones mentioned, expands the points of the solution failure. Nevertheless, the following research will be able to prospect alternatives for these issues, through advanced techniques for domain knowledge integration, improved modality alignment, and fusion approaches. Enhancing dataset consistency and optimizing model architectures becomes crucial to overcoming current limitations, enabling robustness and ethical use of affective computing with multimodalities to obtain a deeper understanding of marketing applications.

Chapter 7

Conclusion

In conclusion, this dissertation 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 affective analysis and the emerging field of multimodality, focusing on combining text, images, audio, and other signals.

The research methodology employed two approaches: the application of the PRISMA methodology for the systematic review of scientific contributions and the CRISP-DM methodology for documenting the phases of the project. The methodologies provided a structured framework for managing the complexities of multimodal data and empower continuous refinement of the model development and performance. The major outcomes were generated by applying AI models that supported multimodal information from two datasets: the E-mail Campaign Dataset, a real-world marketing interactive dataset with customer behavior reports, and the MOSEI, integrating the multimodality of speakers opinions into the analysis of emotional expressions. The developed work revealed that Deep Learning model, such the developed Multimodal Neural Network (MultiNN), significantly outperformed other Machine Learning models, by tackling the entanglement of data into the learning process. The captured affective cues, obtained with the modalities combination, achieved accurate predictions on consumer emotional state and engagement, offering valuable insights to personalize future Marketing strategies and applications.

Nonetheless, some limitations were pinpointed during the research procedures, acknowledging integration complexity during the domain appliance. Additionally, the availability of multimodal datasets in real-world marketing is still challenging, affecting the need to generalize AI tools for real-time marketing functionalities across different communication channels. Another relevant constraint mentioned the ethical and regulation considerations inherent to the problem statement. Due to the rapidly changing perspective of the world's nations on AI, the practice of regulatory guidelines becomes challenging, inhibiting to establish a common knowledge base on permitted and prohibited practices. The same limitations are expected to be solved, allowing all scientific contributions to become usefull, fairly and unbiasedly interventions on the real world.

Bibliography

[1] Ivo Pereira. *Enhancing Phygital Marketing through Multimodal Artificial Intelligence*. 2023. doi: http://doi.org/10.54499/2022.01303.PTDC.

[2] Camille Grange, Izak Benbasat, and Andrew Burton-Jones. "A network-based conceptualization of social commerce and social commerce value". In: *Computers in Human Behavior* 108 (2020), p. 105855. issn: 0747-5632. doi: https://doi.org/10.1016/j.chb.2018.12.033.

[3] Carmen Constantinescu, Bastian Pokorni, and Johannes Wimmer. "Affective Production Systems: Foundations, Reference Model and Roadmap for Implementation and Validation". In: *Procedia CIRP* 104 (2021). 54th CIRP CMS 2021 - Towards Digitalized Manufacturing 4.0, pp. 1783–1786. issn: 2212-8271. doi: https://doi.org/10.1016/j.procir.2021.11.300.

[4] Lindsay McShane et al. "Emoji, Playfulness, and Brand Engagement on Twitter". In: *Journal of Interactive Marketing* 53 (2021), pp. 96–110. issn: 1094-9968. doi: https://doi.org/10.1016/j.intmar.2020.06.002.

[5] Pasquale Del Vecchio, Giustina Secundo, and Antonello Garzoni. "Phygital technologies and environments for breakthrough innovation in customers' and citizens' journey. A critical literature review and future agenda". In: *Technological Forecasting and Social Change* 189 (2023), p. 122342. issn: 0040-1625. doi: https://doi.org/10.1016/j.techfore.2023.122342.

[6] Qi Deng et al. "Speak to head and heart: The effects of linguistic features on B2B brand engagement on social media". In: *Industrial Marketing Management* 99 (2021), pp. 1–15. issn: 0019-8501. doi: https://doi.org/10.1016/j.indmarman.2021.09.005.

[7] Jesus Serrano-Guerrero, Francisco P. Romero, and Jose A. Olivas. "Fuzzy logic applied to opinion mining: A review". In: *Knowledge-Based Systems* 222 (2021), p. 107018. issn: 0950-7051. doi: https://doi.org/10.1016/j.knosys.2021.107018.

[8] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. "A comprehensive survey on sentiment analysis: Approaches, challenges and trends". In: *Knowledge-Based Systems* 226 (2021), p. 107134. issn: 0950-7051. doi: https://doi.org/10.1016/j.knosys.2021.107134.

[9] Jianhua Zhang et al. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Information Fusion* 59 (2020), pp. 103–126. issn: 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.01.011.

[10] Anna Borawska and Małgorzata Łatuszyńska. "The use of neurophysiological measures in studying social advertising effectiveness". In: *Procedia Computer Science* 176 (2020). Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020, pp. 2487–2496. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.09.327.

[11] Rijul Chaturvedi et al. "Social companionship with artificial intelligence: Recent trends and future avenues". In: *Technological Forecasting and Social Change* 193 (2023),

p. 122634. issn: 0040-1625. doi: https://doi.org/10.1016/j.techfore.2023.122634.

[12] Dwi Wahyu Prabowo et al. "A systematic literature review of emotion recognition using EEG signals". In: *Cognitive Systems Research* 82 (2023), p. 101152. issn: 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2023.101152.

[13] Ankita Gandhi et al. "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions". In: *Information Fusion* 91 (2023), pp. 424–444. issn: 1566-2535. doi: https://doi.org/10.1016/j.inffus.2022.09.025.

[14] Wei Li, Zhen Zhang, and Aiguo Song. "Physiological-signal-based emotion recognition: An odyssey from methodology to philosophy". In: *Measurement* 172 (2021), p. 108747. issn: 0263-2241. doi: https://doi.org/10.1016/j.measurement.2020.108747.

[15] Inês César et al. "Multimodal Learning Applications on Digital Marketing: A Review". In: (2023). 23rd International Conference on Hybrid Intelligent Systems (HIS 2023), pp. 1–10. issn: 0263-2241.

[16] Inês César et al. "Exploring multimodal learning applications in marketing: A critical perspective". In: *International Journal of Hybrid Intelligent Systems* Preprint (2024). Preprint, pp. 1–18. issn: 1875-8819. doi: 10.3233/HIS-240018. url: https://doi.org/10.3233/HIS-240018.

[17] Pragya Singh Tomar, Kirti Mathur, and Ugrasen Suman. "Unimodal approaches for emotion recognition: A systematic review". In: *Cognitive Systems Research* 77 (2023), pp. 94–109. issn: 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2022.10.012.

[18] Inês César et al. "A Systematic Review on Responsible Multimodal Sentiment Analysis in Marketing Applications". In: *IEEE Access* 12 (2024), pp. 111943–111961. doi: 10.1109/ACCESS.2024.3441514.

[19] Shaundra B. Daily et al. "Chapter 9 - Affective Computing: Historical Foundations, Current Applications, and Future Trends". In: *Emotions and Affect in Human Factors and Human-Computer Interaction*. Ed. by Myounghoon Jeon. San Diego: Academic Press, 2017, pp. 213–231. isbn: 978-0-12-801851-4. doi: https://doi.org/10.1016/B978-0-12-801851-4.00009-4. url: https://www.sciencedirect.com/science/article/pii/B9780128018514000094.

[20] Resham Arya, Jaiteg Singh, and Ashok Kumar. "A survey of multidisciplinary domains contributing to affective computing". In: *Computer Science Review* 40 (2021), p. 100399. issn: 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2021.100399. url: https://www.sciencedirect.com/science/article/pii/S1574013721000393.

[21] Nusrat J. Shoumy et al. "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals". In: *Journal of Network and Computer Applications* 149 (2020), p. 102447. issn: 1084-8045. doi: https://doi.org/10.1016/j.jnca.2019.102447. url: https://www.sciencedirect.com/science/article/pii/S1084804519303078.

[22] Sze Chit Leong et al. "Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing". In: *Computer Science Review* 48 (2023), p. 100545. issn: 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2023.100545. url: https://www.sciencedirect.com/science/article/pii/S1574013723000126.

[23] Rajat Kumar Behera et al. "Cognitive computing based ethical principles for improving organisational reputation: A B2B digital marketing perspective". In: *Journal of Business*

Research 141 (2022), pp. 685–701. issn: 0148-2963. doi: https://doi.org/10.1016/j.jbusres.2021.11.070. url: https://www.sciencedirect.com/science/article/pii/S0148296321008778.

[24] Liye Ma and Baohong Sun. "Machine learning and AI in marketing – Connecting computing power to human insights". In: *International Journal of Research in Marketing* 37.3 (2020), pp. 481–504. issn: 0167-8116. doi: https://doi.org/10.1016/j.ijresmar.2020.04.005. url: https://www.sciencedirect.com/science/article/pii/S0167811620300410.

[25] Sergei Polevikov. "Advancing AI in Healthcare: A Comprehensive Review of Best Practices". In: *Clinica Chimica Acta* 548 (Aug. 2023), p. 117519. doi: 10.1016/j.cca.2023.117519.

[26] István Mezgár and József Váncza. "From ethics to standards – A path via responsible AI to cyber-physical production systems". In: *Annual Reviews in Control* 53 (2022), pp. 391–404. issn: 1367-5788. doi: https://doi.org/10.1016/j.arcontrol.2022.04.002. url: https://www.sciencedirect.com/science/article/pii/S1367578822000177.

[27] Shahab Saquib Sohail et al. "Decoding ChatGPT: A taxonomy of existing research, current challenges, and possible future directions". In: *Journal of King Saud University - Computer and Information Sciences* 35.8 (2023), p. 101675. issn: 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2023.101675. url: https://www.sciencedirect.com/science/article/pii/S131915782300229X.

[28] Rostam J. Neuwirth. "Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA)". In: *Computer Law & Security Review* 48 (2023), p. 105798. issn: 0267-3649. doi: https://doi.org/10.1016/j.clsr.2023.105798. url: https://www.sciencedirect.com/science/article/pii/S0267364923000092.

[29] Natalia Díaz-Rodríguez et al. "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation". In: *Information Fusion* 99 (2023), p. 101896. issn: 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101896.

[30] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443. doi: 10.1109/TPAMI.2018.2798607.

[31] Catrin Sohrabi et al. "PRISMA 2020 statement: What's new and the importance of reporting guidelines". In: *International Journal of Surgery* 88 (2021), p. 105918. issn: 1743-9191. doi: https://doi.org/10.1016/j.ijsu.2021.105918.

[32] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. "A Systematic Literature Review on Applying CRISP-DM Process Model". In: *Procedia Computer Science* 181 (2021). CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, pp. 526–534. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.01.199.

[33] William Gu et al. "Informational vs. emotional B2B firm-generated-content on social media engagement: Computerized visual and textual content analysis". In: *Industrial Marketing Management* 112 (2023), pp. 98–112. issn: 0019-8501. doi: https://doi.org/10.1016/j.indmarman.2023.04.012.

[34] Ming-Hui Huang and Roland T. Rust. "A Framework for Collaborative Artificial Intelligence in Marketing". In: *Journal of Retailing* 98.2 (2022), pp. 209–223. issn: 0022-4359. doi: https://doi.org/10.1016/j.jretai.2021.03.001.

[35] Wei Xu et al. "How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce". In: *Decision Support Systems* 172 (2023), p. 113984. issn: 0167-9236. doi: https://doi.org/10.1016/j.dss.2023.113984.

[36] Abolfazl Mehbodniya et al. "Online product sentiment analysis using random evolutionary whale optimization algorithm and deep belief network". In: *Pattern Recognition Letters* 159 (2022), pp. 1–8. issn: 0167-8655. doi: https://doi.org/10.1016/j.patrec.2022.04.024.

[37] Tsun-hin Cheung and Kin-man Lam. "Crossmodal bipolar attention for multimodal classification on social media". In: *Neurocomputing* 514 (2022), pp. 1–12. issn: 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.09.140.

[38] Farzaneh Jouyandeh and Pooya Moradian Zadeh. "IPARS: An Image-based Personalized Advertisement Recommendation System on Social Networks". In: *Procedia Computer Science* 201 (2022). The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40), pp. 375–382. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2022.03.050.

[39] Akshi Kumar et al. "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data". In: *Information Processing & Management* 57.1 (2020), p. 102141. issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102141.

[40] Luna De Bruyne et al. "Aspect-Based Emotion Analysis and Multimodal Coreference: A Case Study of Customer Comments on Adidas Instagram Posts". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 574–580. url: https://aclanthology.org/2022.lrec-1.61.

[41] Donghong Gu et al. "Targeted Aspect-Based Multimodal Sentiment Analysis: An Attention Capsule Extraction and Multi-Head Fusion Network". In: *IEEE Access* 9 (2021), pp. 157329–157336. issn: 2169-3536. doi: 10.1109/ACCESS.2021.3126782.

[42] Mohina Gandhi and Arpan Kumar Kar. "How do Fortune firms build a social presence on social media platforms? Insights from multi-modal analytics". In: *Technological Forecasting and Social Change* 182 (2022), p. 121829. issn: 0040-1625. doi: https://doi.org/10.1016/j.techfore.2022.121829.

[43] Vasco Lopes et al. "An AutoML-based Approach to Multimodal Image Sentiment Analysis". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2021. doi: 10.1109/ijcnn52387.2021.9533552.

[44] Guangxia Xu, Weifeng Li, and Jun Liu. "A social emotion classification approach using multi-model fusion". In: *Future Generation Computer Systems* 102 (2020), pp. 347–356. issn: 0167-739X. doi: https://doi.org/10.1016/j.future.2019.07.007.

[45] Han Lin et al. "PS-Mixer: A Polar-Vector and Strength-Vector Mixer Model for Multimodal Sentiment Analysis". In: *Information Processing & Management* 60.2 (2023), p. 103229. issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.103229.

[46] Jie Zhou et al. "MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis". In: *Neurocomputing* 455 (2021), pp. 47–58. issn: 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.05.040.

[47] Hengyun Li et al. "Is a picture worth a thousand words? Understanding the role of review photo sentiment and text-photo sentiment disparity using deep learning algorithms". In: *Tourism Management* 92 (2022), p. 104559. issn: 0261-5177. doi: https://doi.org/10.1016/j.tourman.2022.104559.

[48] Siyu Zhu et al. "A new approach for product evaluation based on integration of EEG and eye-tracking". In: *Advanced Engineering Informatics* 52 (2022), p. 101601. issn: 1474-0346. doi: https://doi.org/10.1016/j.aei.2022.101601.

[49] Yuntao Shou et al. "Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis". In: *Neurocomputing* 501 (2022), pp. 629–639. issn: 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.06.072.

[50] Yazhou Zhang et al. "A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis". In: *Information Fusion* 62 (2020), pp. 14–31. issn: 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.04.003.

[51] Jyotirmoy Karjee et al. "A Lightweight Multimodal Learning Model to Recognize User Sentiment in Mobile Devices". In: *2023 IEEE International Conference on Consumer Electronics (ICCE)*. Jan. 2023, pp. 1–6. doi: 10.1109/ICCE56470.2023.10043524.

[52] Yang Qian et al. "Popularity prediction for marketer-generated content: A text-guided attention neural network for multi-modal feature fusion". In: *Information Processing & Management* 59.4 (2022), p. 102984. issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.102984.

[53] Dinghao Xi et al. "A multimodal time-series method for gifting prediction in live streaming platforms". In: *Information Processing & Management* 60.3 (2023), p. 103254. issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2022.103254.

[54] Wei Bi et al. "Enterprise Strategic Management From the Perspective of Business Ecosystem Construction Based on Multimodal Emotion Recognition". In: *Frontiers in Psychology* 13 (2022). issn: 1664-1078. doi: 10.3389/fpsyg.2022.857891.

[55] Kyeonghun Kim and Sanghyun Park. "AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis". In: *Information Fusion* 92 (2023), pp. 37–45. issn: 1566-2535. doi: https://doi.org/10.1016/j.inffus.2022.11.022.

[56] Dhruv Grewal et al. "The Future of Digital Communication Research: Considering Dynamics and Multimodality". In: *Journal of Retailing* 98.2 (2022), pp. 224–240. issn: 0022-4359. doi: https://doi.org/10.1016/j.jretai.2021.01.007.

[57] Zheng Wang, Peng Gao, and Xuening Chu. "Sentiment analysis from Customer-generated online videos on product review using topic modeling and Multi-attention BLSTM". In: *Advanced Engineering Informatics* 52 (2022), p. 101588. issn: 1474-0346. doi: https://doi.org/10.1016/j.aei.2022.101588.

[58] Jinyu Chen et al. "The Multimodal Emotion Information Analysis of E-Commerce Online Pricing in Electronic Word of Mouth". In: *Journal of Global Information Management* 30 (Jan. 2022), pp. 1–17. doi: 10.4018/JGIM.315322.

[59] Luwei Xiao et al. "Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis". In: *Information Processing & Management* 60.6 (2023), p. 103508. issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2023.103508.

[60] Sana Rahmani et al. "Transfer-based adaptive tree for multimodal sentiment analysis based on user latent aspects". In: *Knowledge-Based Systems* 261 (2023), p. 110219. issn: 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.110219.

[61] Ayush Kumar and Jithendra Vepa. "Gated Mechanism for Attention Based Multi Modal Sentiment Analysis". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2020, pp. 4477–4481. doi: 10.1109/ICASSP40776.2020.9053012.

[62] Meng Xu et al. "CMJRT: Cross-Modal Joint Representation Transformer for Multimodal Sentiment Analysis". In: *IEEE Access* 10 (2022), pp. 131671–131679. issn: 2169-3536. doi: 10.1109/ACCESS.2022.3219200.

[63] Ajwa Aslam, Allah Bux Sargano, and Zulfiqar Habib. "Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks". In: *Applied Soft Computing* 144 (2023), p. 110494. issn: 1568-4946. doi: https://doi.org/10.1016/j.asoc.2023.110494.

[64] Qun Wu et al. "Emotion classification on eye-tracking and electroencephalograph fused signals employing deep gradient neural networks". In: *Applied Soft Computing* 110 (2021), p. 107752. issn: 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.107752.

[65] Alexandru Capatina et al. "Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations". In: *Technological Forecasting and Social Change* 151 (2020), p. 119794. issn: 0040-1625. doi: https://doi.org/10.1016/j.techfore.2019.119794.

[66] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. "BERTimbau: Pretrained BERT Models for Brazilian Portuguese". In: *Intelligent Systems*. Ed. by Ricardo Cerri and Ronaldo C. Prati. Cham: Springer International Publishing, 2020, pp. 403–417. isbn: 978-3-030-61377-8.

[67] MultiComp Lab. *CMU-MOSEI Dataset*. 2018. doi: http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/.