An Evaluation of Multi-Modal Approaches for Sentiment Analysis using Deep Learning

Kamal Kamal

MASTER'S THESIS

A research components submitted to Auckland University of Technology in fulfilment of the requirements for the degree of

Master of Philosophy (M.Phil)

Supervisor

Dr Sira Yongchareon

School of Engineering, Computer and Mathematical Sciences

Auckland,August 2024

"The only way to do great work is to love what you do." – **Steve Jobs**

Attestation of Authorship

I affirm that this submission results from my independent effort. To the best of my understanding and belief, it does not include any content that has been previously published or authored by another individual, except as explicitly acknowledged. Furthermore, it does not incorporate material that has been substantially used in seeking any other academic degree or diploma from any university or institution of higher learning.

28 May 2024 Kamal Kamal

Acknowledgements

This thesis represents a segment of the Master of Philosophy, Computer and Mathematical Sciences, Auckland University of Technology, New Zealand. I extend my appreciation to my parents for their steadfast support and to my friends for their encouragement throughout my academic journey in New Zealand. Special gratitude is reserved for my supervisor, Dr. Sira Yongchareon, whose invaluable assistance significantly enriched my learning experience. Dr. Yongchareon initially outlined a comprehensive plan for my research project, and over the subsequent year, we held regular meetings to discuss and organize tasks for each week. Under his guidance, I successfully completed my thesis within the specified timeframe, with Dr. Yongchareon provides patient guidance for problem-solving at every step.

Auckland, August 2024

Kamal

Abstract

This thesis presents an evaluation of eight deep-learning-based multimodal models for sentiment analysis, applied to the CMU Multimodal Opinion Sentiment Expression in Recordings (CMU MOSEI) and CMU Multimodal Sentiment Opinions (CMU MOSI) benchmark datasets. The study delves into the realm of multimodal sentiment analysis by integrating text, audio, and visual data. Through rigorous experimentation and analysis, key observations are drawn regarding the performance and effectiveness of each model. To evaluate the models' performance in classifying positive and negative sentiment across various modalities, the study employs a range of metrics including accuracy, F1-score, loss, mean absolute error (MAE), and correlation. Beyond demonstrating the promise of multimodal sentiment analysis, this research offers valuable knowledge. It sheds light on both the optimal configurations for models and how incorporating various modalities impacts the accuracy of sentiment classification.

Contents

List of Figures

List of Tables

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BPTT	Backpropagation Through Time
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
LMF	Low-rank Multimodal Fusion
LSTM	Long Short-Term Memory Networks
MAE	Mean Absolute Error
MI	Mutual Information
MISA	Multimodal Invariant and Modality Specific Representations for Multimodal Sentiment Analysis
MOSEI	Multimodal Opinion Sentiment and Emotion Intensity
MOSI	Multimodal Corpus of Sentiment Intensity
MSA	Multimodal Sentiment Analysis
NLP	Natural Language Processing
PCA	Principal Component Analysis
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
SVD	Singular Value Decomposition
TETFN	Text Enhanced Transformer Fusion Network

"Education is the passport to the future, for tomorrow belongs to those who prepare for it today."

– Malcolm X

CHAPTER 1

Introduction

The proliferation of social media and the widespread adoption of smartphones have elevated videos to a prominent platform for articulating viewpoints; and disseminating product reviews. Robust user-generated content on platforms such as YouTube, Facebook, and Twitter underscores the imperative for advanced sentiment analysis to decipher the intricacies of expressed opinions. One primary factor contributing to this phenomenon is the immense volume of data produced daily across different mediums like text, audio, and video on social media platforms worldwide [5] [6]. Sentiment analysis finds extensive application across diverse sectors including business, government, biomedicine, and recommender systems. It involves categorizing opinions into positive, negative, or neutral sentiments [7], a pivotal task involving the discernment of the emotional tenor or attitude manifested in a given piece of content, assumes paramount importance in comprehending user sentiments, preferences, and reactions across various domains such as social media, e-commerce, and customer reviews.

In the realm of market research, sentiment analysis plays a crucial role. By analyzing customer feedback, companies can gain valuable insights into customer satisfaction and brand perception and social media [8]. Political parties use it to understand public opinion, aiding in campaign strategy[9]; Automatic sentiment analysis, powered by NLP, categorizes text to provide efficient insights into opinions on various topics [10].

The amalgamation of multiple modalities, notably the fusion of textual and visual elements, facilitates a more exhaustive comprehension of sentiment. This is particularly germane, given that human communication inherently integrates a melange of verbal and non-verbal cues;

emotions expressed by users transcend mere linguistic articulation, extending into the realm of visual cues, tonal inflections, and diverse modalities. Consequently, the systematic analysis of multi-modal content emerges as an indispensable imperative for attaining a nuanced and precision-laden portrayal of sentiment dynamics within real-world scenarios.

In the field of Natural Language Processing (NLP), sentiment analysis, or opinion mining, aims to identify the emotional tone conveyed within a text. The principal aim is to perform sentiment polarity classification, categorizing the expressed opinions as positive, negative, or neutral. From gauging public opinion on social media to assessing customer satisfaction through feedback analysis, sentiment analysis is a powerful tool for extracting emotional sentiment from textual data across diverse applications.

Sentiment Analysis is the task of classifying the polarity of a given text/image/audio. For instance, a text-based tweet can be categorized into either "positive," "negative," or "neutral". Given the text and accompanying labels, a model can be trained to predict the correct sentiment.

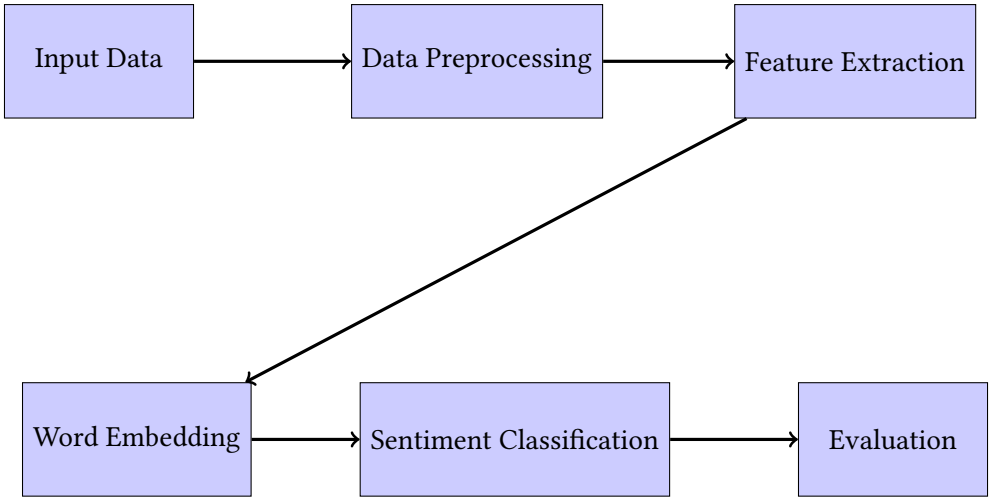

Figure 1.1: Steps in Sentiment Analysis

1.1 illustrates the sequential stages of sentiment analysis. The input data, comprising textual sentiments, undergoes text preprocessing steps including stop word removal, tokenization, stemming, and conversion to lowercase. Subsequently, the processed text is converted into vectors through techniques like word2vec and GloVe. Various deep learning architectures such as Convolutional Neural Network (CNN)s, LSTMs, and Bi-LSTMs and also Recurrent Neural Network (RNN)s are frequently employed for sentiment analysis. Performance assessment can be conducted using diverse metrics including accuracy, precision, recall, Loss, and F-score.

1.0.1 Verification of Theoretical Frameworks

Within the realm of academic inquiry into multimodal sentiment analysis, there is frequent engagement in crafting innovative theoretical frameworks, methodologies, and models. The validation process functions as a means of validating, systematically examining the effectiveness and relevance of these theoretical constructs through empirical testing processes. Through methodical evaluations, researchers can substantiate or enhance their hypotheses, contributing to the progression of the theoretical foundations of multimodal sentiment analysis.

1.0.2 Assessing Model Performance through Benchmarking

A pivotal element of the evaluation procedure involves benchmarking the performance of models in multimodal sentiment analysis. Through subjecting these models to standardized evaluations, researchers set a benchmark for comparative analysis. This benchmark not only facilitates the assessment of the efficacy of current models but also serves as a reference for future innovations. It becomes a foundational point for monitoring the progression of multimodal sentiment analysis techniques over time.

1.0.3 Unveiling Model Limitations

The evaluation process elucidates the constraints and difficulties encountered by extant multimodal sentiment analysis models. Through a systematic examination of model performance across diverse dimensions, researchers acquire discernment into the circumstances in which these models may exhibit deficiencies. Grasping these limitations is imperative for honing current models and guiding forthcoming research endeavors to tackle particular challenges.

1.0.4 Recommendations for Model Selection in Practical Applications

The practical utilization of multimodal sentiment analysis in domains like marketing, customer service, and social media analytics is directly enhanced through the evaluation process. The results offer insights into choosing the most efficacious models for real-world situations. Businesses and industries can make judicious decisions regarding the implementation of sentiment analysis tools, aligning them with their particular needs and objectives.

1.0.5 Recognition of Research Gaps

The evaluation process frequently reveals domains in which current approaches to multimodal sentiment analysis may be deficient or inadequately explored. These identified gaps in knowledge serve as stimuli for subsequent research, encouraging scholarly exploration into specific facets of multimodal sentiment analysis that necessitate improvement, innovation, or entirely novel methodologies.

1.0.6 Generalizability Enhancement

Rigorous evaluations are crucial for strengthening the adaptability (generalizability) of multimodal sentiment analysis models. By testing these models on diverse datasets and application areas, researchers can refine them to be more robust and universally applicable. This focus on generalizability ensures that advancements in the field translate into real-world benefits across a wider range of contexts.

Fundamentally, the assessment of multimodal sentiment analysis functions as a dynamic conduit, seamlessly connecting theoretical progressions with pragmatic applications. This evaluative process goes beyond merely validating academic theories; it extends its impact to furnish practical insights and guidelines tailored for industry professionals. This symbiotic relationship between theory and practice contributes significantly to the evolution and proficient deployment of multimodal sentiment analysis within real-world contexts.

In essence, the evaluation serves as a reciprocal bridge: it ensures that theoretical advancements are not confined to abstract academic discourse but are instead translated into actionable insights that industry practitioners can leverage. The practical guidelines derived from this evaluation empower professionals to navigate the nuanced landscape of multimodal sentiment analysis, fostering its maturation and optimizing its efficacy when applied in diverse real-world scenarios. This iterative cycle of evaluation and application facilitates a continuous feedback loop, refining both theoretical frameworks and practical methodologies, ultimately enhancing the overall competence of multimodal sentiment analysis in addressing the multifaceted challenges posed by contemporary communication modalities.

1.0.7 Sentiment Analysis definition on Different Modalities

- **Text:** Comprises written material, including articles, captions, comments, and descriptions.
- **Audio:** Encompasses spoken language, tone, pitch, and other auditory features derived from sources such as podcasts, interviews, or videos.
- **Video:** Refers to visual content that includes facial expressions, gestures, and contextual visual details from videos.
- **Images:** Denotes static visual content, often paired with descriptions or captions that offer contextual understanding.

1.1 Research Questions and Contributions

The central focus of this thesis revolves around evaluating the performance and efficacy of eight implemented multimodal sentiment analysis models using the Multimodal Corpus of Sentiment Intensity (MOSI) and Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) datasets. The research questions guiding this study are structured to delve into the nuances of model evaluation and comparison, as outlined below:

RQ1: Which multimodal sentiment analysis models achieve the highest performance across critical evaluation metrics such as accuracy, precision, recall, and F1-score when applied to the challenging sentiment analysis tasks posed by the MOSI and MOSEI datasets? What insights can be derived from their comparative analysis based on these metrics?

RQ2: What are the respective strengths and weaknesses of each multimodal sentiment analysis model in effectively capturing and interpreting nuanced sentiments across diverse modalities, including text, audio, and visual cues, measured in terms of precision, recall, F1-score, and accuracy?

RQ3: How can the performance of the multimodal sentiment analysis models be further optimized or enhanced, and what novel strategies or modifications can be proposed based on the evaluation results, particularly in the context of precision, recall, F1-score, and accuracy?

By addressing these research questions, the thesis aims to conduct a comprehensive and rigorous evaluation of the implemented multimodal sentiment analysis models, emphasizing

the significance of precision, recall, F1-score, and accuracy as key performance indicators. Through statistical analysis and comparative assessments, this research contributes valuable insights and advancements to the domain of multimodal sentiment analysis methodology and application, specifically focusing on precision, recall, F1-score, and accuracy metrics to measure model performance and effectiveness.

1.2 Thesis Structure

The structure of this thesis adheres to a rigorous academic framework, comprising five primary chapters that collectively contribute to the investigation and analysis of multimodal sentiment analysis. Chapter 2, "Related Work," meticulously surveys the existing literature on sentiment analysis, covering a spectrum of modalities such as text, image, audio, and their combinations in multimodal contexts. This chapter not only synthesizes the diverse methodologies employed in sentiment analysis but also evaluates the efficacy and limitations of current multimodal sentiment analysis approaches. Moreover, it examines the implementation and results of eight distinct multimodal sentiment analysis models, offering a comprehensive overview of their performance metrics and comparative analyses. Chapter 3, "Methodology," articulates the research methodology adopted in this study, elucidating the selection, preprocessing, and characteristics of the datasets utilized. It also delineates the evaluation metrics employed for rigorously assessing the performance and efficacy of the implemented multimodal sentiment analysis models. Chapter 4, "Results and Discussions," presents the findings derived from the implementation of the multimodal sentiment analysis models and engages in scholarly discourse to interpret, analyze, and contextualize these results within the broader landscape of sentiment analysis research. The chapter explores a nuanced discussion on the implications, strengths, limitations, and potential future directions of multimodal sentiment analysis, substantiated with visual aids and comparative analyses to augment clarity and scholarly rigor. Lastly, Chapter 5, "Conclusion and Future Scope," This work culminates by synthesizing the study's key takeaways and charting potential avenues for future exploration and advancements in multimodal sentiment analysis. This chapter offers a reflective analysis of the research outcomes, discusses the broader implications of the findings, and proposes potential directions for enhancing and advancing multimodal sentiment analysis

methodologies. Through this structured academic approach, the thesis endeavors to contribute meaningfully to the ongoing discourse and advancement of multimodal sentiment analysis methodologies.

CHAPTER 2

Related Work

In the realm of multimodal sentiment analysis, the integration of diverse modalities plays a pivotal role in capturing the richness and complexity of human emotions. Analyzing sentiment from textual data has been a cornerstone of sentiment analysis research. Numerous studies have explored NLP techniques to achieve this goal. Approaches such as sentiment lexicons, machine learning models, and deep learning architectures have been employed to decipher the emotional undertones conveyed through written expressions.

Simultaneously, the audio modality has garnered attention as a valuable source of emotional cues. Research in this domain involves analyzing acoustic features, intonation, and speech patterns to infer sentiment and emotional states. Techniques ranging from signal processing techniques to deep learning-based audio analysis have been instrumental in decoding the intricate relationship between acoustic signals and emotional content.

Expanding beyond textual and auditory dimensions, the incorporation of visual modalities, such as images, has emerged as a critical facet in multimodal sentiment analysis. Visual cues, including facial expressions, body language, and scene context, offer additional layers of information that contribute significantly to understanding sentiment. Researchers have explored computer vision techniques, including facial emotion recognition algorithms and image-based deep learning models, to extract sentiment-related features from visual content.

The synergy of text, audio, and image modalities has driven a paradigm shift towards comprehensive multimodal sentiment analysis, where the fusion of information from different sources enriches the understanding of complex emotional states and enhances the overall

accuracy and depth of sentiment analysis systems.

Moreover, the convergence of these modalities has given rise to sophisticated fusion strategies, where collective information from text, audio, and image sources is harnessed to achieve a more holistic sentiment understanding. Fusion techniques range from early fusion, combining modalities at the feature level, to late fusion, integrating modality-specific outputs at a higher level. This integration not only addresses the limitations inherent in individual modalities but also unlocks new possibilities for uncovering subtle nuances in sentiment that may be missed when analyzing each modality in isolation.

The exploration of multimodal sentiment analysis has extended beyond traditional datasets, with researchers increasingly incorporating diverse datasets that encompass real-world scenarios, such as social media interactions, video commentaries, and online reviews. This shift towards more ecologically valid datasets reflects the growing interest in understanding sentiment in complex, dynamic, and naturalistic settings, where multiple modalities synergistically contribute to the intricate tapestry of human emotional expression.

As the field progresses, challenges related to multimodal sentiment analysis continue to motivate advancements in methodologies and techniques. Questions surrounding the optimal fusion strategies, feature representation across modalities, and scalability of models to accommodate the growing complexity of multimodal data are areas of ongoing investigation. Additionally, recent endeavors have delved into the ethical considerations associated with the use of multimodal sentiment analysis, particularly concerning privacy concerns related to the analysis of audio-visual content. This intersection of technology, emotion, and ethics underscores the importance of developing robust and responsible multimodal sentiment analysis approaches that not only enhance our understanding of sentiment but also align with ethical considerations in the evolving landscape of information technology.

Literature Review Structure. The literature review presented in this thesis is structured to provide a thorough understanding of the research landscape, particularly concerning the use of different modalities in sentiment analysis and the role of the MOSI and MOSEI datasets. This review aims to explore the foundational theories and concepts, followed by an in-depth analysis of contemporary research, which focuses on the functionality and applications of these datasets.

The review begins by discussing the foundational theories in multimodal sentiment analysis,

establishing a framework for understanding how different modalities are integrated and analyzed. Following this, it delves into the specifics of the MOSI and MOSEI datasets, critically examining their development, key features, and how they have been utilized in recent studies. This approach not only highlights the evolution of research in this area but also identifies the challenges and opportunities that have emerged.

By synthesizing these insights, the literature review sets the stage for the research questions and objectives that will be addressed in the subsequent sections of this thesis.

2.1 Sentiment Analysis on Different Modalities

NLP encompasses various tasks, with sentiment analysis playing a key role in extracting emotional sentiment from text data, has traditionally focused on textual data. However, with the proliferation of multimedia content, sentiment analysis has expanded to encompass various modalities such as images, videos, and audio. This chapter delves into the exploration of sentiment analysis across different modalities, aiming to uncover insights and challenges associated with analyzing sentiment in non-textual data.

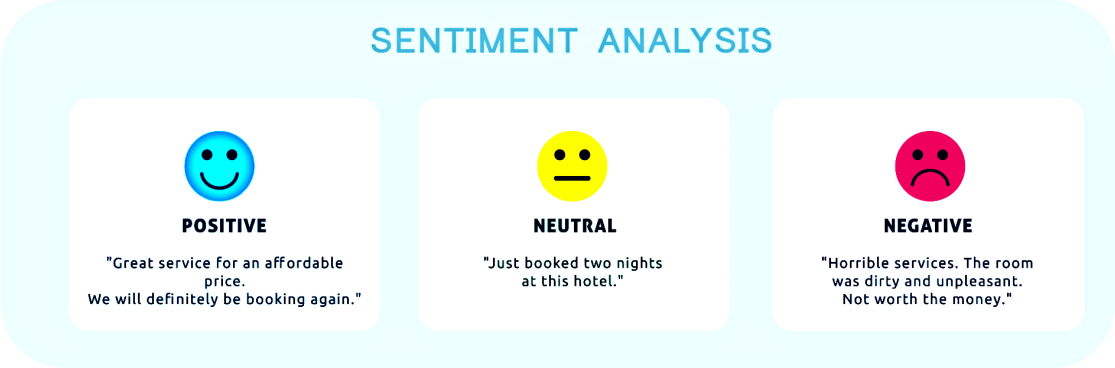

Figure 2.1: Classification of Text Sentiment Analysis on customer reviews

2.1.1 Text Sentiment Analysis

Text Sentiment Analysis, or sentiment analysis, represents an intricate lineage tracing back to the pre-2000s, originating with lexical-based approaches. These early frameworks predicated sentiment determination on the analysis of individual lexical units and their corresponding affective connotations. This rudimentary paradigm laid the foundational framework, serving

as the genesis for subsequent methodological progressions. As shown in Figure 2.1, the classification of text sentiment analysis on customer reviews highlights the effectiveness of various models.

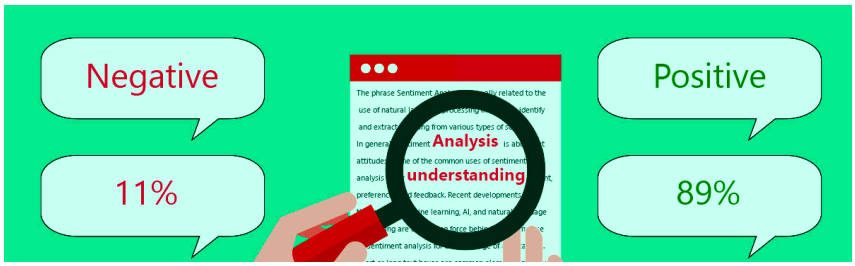

Figure 2.2: Textual Sentiment Analysis

During the 1990s, the development of sentiment lexicons became prominent. These lexicons were essentially lists of words associated with positive or negative sentiment. Researchers compiled these lists to aid in sentiment analysis by assigning a sentiment score to words, allowing for a more systematic approach.

There are ways to pick important words when looking at topics or sentiments in text. In 2002, a novel approach proposed the effectiveness of applying machine learning techniques to the sentiment classification problem[11]. This method works well for general text sorting, showing that just looking at how often a word appears can be a good way to decide which words are important for sentiment analysis.

Lexicon-based[12] methodologies initiate with a limited set of initial 'seed' words and employ processes such as synonym detection or utilization of diverse online resources to iteratively expand this set, thereby acquiring a more extensive lexicon.

Numerous sentiment lexicons have become notable in both sentiment analysis research and practical applications. One instance is the AFINN-111[13] Lexicon, which provides numerical scores to words indicating their sentiment intensity. Another extensively utilized resource is the SentiWordNet[14] lexicon, associating terms with synsets and assigning sentiment scores. These lexicons serve as essential tools for researchers, frequently employed as foundational resources for benchmarking and constructing sentiment analysis models.

The utilization of sentiment lexicons underscores the interdisciplinary nature of sentiment analysis, amalgamating linguistics, computer science, and cognitive science to unravel the intricate tapestry of human emotions as manifested through language. Ongoing research continues to refine and expand sentiment lexicons to enhance their applicability across diverse

linguistic contexts.

Figure 2.2 illustrates the classification of positive and negative sentiments in textual sentiment analysis. It emerged as an alternative to topic detection, with early work proposing automatic detection of directionality[15] and identification of the point of view[16]. Pang and Lee's influential review in 2008 highlighted the growing interest in sentiment analysis[17]. Supervised methods, such as those used in the SemEval competition[18], have been instrumental in text sentiment analysis, with SVM-based models showing promise[19][20][21]. Recent advancements include the use of convolutional neural networks, as demonstrated by SwissCheese[22].

Transitioning into the mid-2000s, a discernible shift manifested towards rule-based systems. These systems leveraged predetermined linguistic rules to deduce sentiment orientation from textual data. The inherent limitations of rule-based methodologies, particularly their incapacity to adeptly navigate contextual intricacies and accommodate diverse language patterns, instigated a quest for more sophisticated analytical paradigms.

The advent of machine learning ushered in a pivotal juncture in the evolution of sentiment analysis during the mid to late 2000s. Supervised learning algorithms, particularly classifiers, ascended to prominence. These models underwent training on annotated datasets, facilitating the discernment of intricate patterns and the formulation of predictive inferences regarding sentiment based on acquired features.

Technological progression witnessed the pervasive integration of NLP techniques. NLP[23], instrumental in achieving a more nuanced comprehension of language by assimilating syntactic and semantic structures, catalyzed a paradigm shift. This transition encompassed the assimilation of sentiment lexicons, word embeddings, and the exploration of sentiment analysis within the ambit of aspect-based sentiment analysis, wherein sentiments affiliated with distinct facets or attributes within a given text were systematically delineated.

The introduction of transformers for text occurred with the paper titled "Attention is All You Need"[24]. The authors introduced the Transformer model, which is a type of neural network architecture that relies on self-attention mechanisms.

Emerging as a cornerstone in the field of NLP, the Transformer architecture has revolutionized various tasks. Its impact is evident in the significant advancements achieved in areas like machine translation, text summarization, and language understanding, propelling

the state-of-the-art. The attention mechanism in transformers allows the model to focus on different parts of the input sequence when making predictions, equipping the model to bridge distant elements and grasp the broader context efficiently.

This paper marked a significant departure from the previously dominant recurrent neural network (RNN) and convolutional neural network (CNN) architectures for sequence modeling, showcasing the effectiveness of transformers in handling sequential data, including natural language.

In the contemporary echelon of development, deep learning models, prominently featuring RNNs and transformer architectures, have ascended to eminence. These models exhibit prowess in capturing intricate dependencies latent within textual data, thereby enabling the discernment of subtle nuances inherent in sentiment expression. The introduction of pre-trained language models, exemplified by BERT[25] (Bidirectional Encoder Representations from Transformers), has fundamentally reshaped sentiment analysis. This paradigm shift facilitates transfer learning, augmenting model generalization across diverse datasets and exemplifying a seminal evolution in the analytical landscape.

Bidirectional Encoder Representations from Transformers (BERT) is a powerful natural language processing model that has revolutionized the field of machine learning and NLP. Developed by Google in 2018, BERT is based on the Transformer architecture, allowing it to capture contextual information from both the left and right sides of a word in a sentence.

Upon its publication, BERT demonstrated leading performance in several natural language understanding tasks. Since its initial release, BERT has undergone several iterations and improvements to enhance its performance and versatility. BERT stands as a sophisticated and advanced language model, facilitating automated language comprehension. Its capability to achieve cutting-edge performance is underpinned by extensive training on vast datasets and harnessing the transformative power of the Transformers architecture, thereby reshaping the landscape of NLP.

The continuous evolution of BERT has spurred groundbreaking research and has become an indispensable tool in industries leveraging NLP technologies. Researchers and practitioners alike are exploring novel ways to harness the capabilities of the latest version to address real-world challenges and push the boundaries of what is achievable in natural language understanding.

In the realm of natural language processing and sentiment analysis, BERT is a powerful pre-trained language model. The layers of BERT capture different levels of linguistic information, ranging from low-level details to more abstract and contextual representations. The study titled 'Methods to Enhance BERT in Aspect-Based Sentiment Classification'[26] emphasizes the significance of utilizing the outputs from layers 9-12 of BERT in sentiment classification tasks. Each layer in BERT captures different aspects of language features and context. Layers closer to the input (lower layers) may capture more basic syntactic and semantic structures, while higher layers capture more complex patterns and context.

By emphasizing the importance of utilizing the outputs from layers 9-12, the paper suggests that these specific layers play a crucial role in understanding and representing sentiment-related information within the input text. This insight could be based on experimental evidence or analysis conducted by the authors, showcasing the specific effectiveness of these layers in enhancing sentiment classification performance.

Supervised sentiment analysis often focuses on specific domains like social media, with Twitter being prominent. SemEval competitions have driven advancements in this field, with SVM models initially leading[19], later enhanced with richer lexicon resources[22]. Recent approaches combine SVM with other methods, showing promising results[21]. Notably, SwissCheese achieved state-of-the-art performance using convolutional neural networks trained on tweet datasets. Alternative benchmarks favor a bootstrap parametric ensemble framework[27][28].

Notwithstanding these strides, challenges endure within the field, encompassing the nuanced interpretation of sarcasm, irony, and contextual variations in sentiment expression. The continual trajectory of Text Sentiment Analysis is propelled by ongoing interdisciplinary collaborations, progressive strides in machine learning methodologies, and the seamless assimilation of state-of-the-art technologies. This dynamic evolution ensures the method's adaptability to the dynamic contours of evolving language patterns, underscoring its persistent relevance across varied applications, including business intelligence, customer feedback analysis, and social media monitoring.

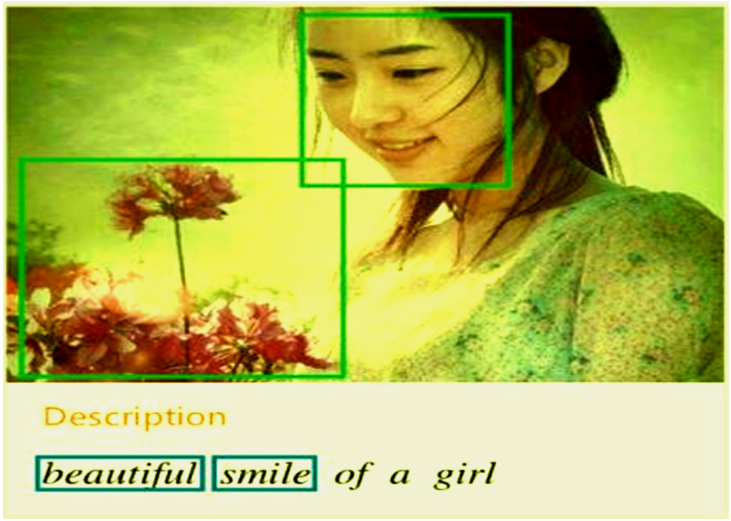

Figure 2.3: Image and Text Sentiment Analysis.

2.1.2 Image Sentiment Analysis

Image Sentiment Analysis is a computational method that employs mathematical techniques to understand and classify emotions expressed in pictures. It involves training models on labeled datasets, optimizing parameters, and assessing performance using metrics. By extracting features from images and mapping them to sentiment labels, this process aids in developing accurate models for understanding emotions conveyed through visual content.

Exploring the realm of visual sentiment analysis is a relatively recent endeavor, distinct from established fields like emotion recognition in vision. This emerging area targets sentiment expression through facial and bodily gestures or in visual multimedia. In particular, research focuses on detecting sentiment from observable expressions and interpreting sentiment associated with visual content.

A study by [29] acknowledged that recommender systems often predict a user's preference for an item, commonly employing matrix factorization for rating estimation [30]. The role of images in e-commerce has also been recognized [31, 32], where they function as additional features. Users have shown a tendency to favor items with similar images [33, 34]. However, sentiment analysis, unlike recommendation systems, evaluates the polarity of an image itself, rather than the user-item relationship. In visually-aware recommendation models, the focus shifts to suggesting products resembling a given photo, resembling an image retrieval task [35].

While the automatic extraction of facial expressions and bodily gestures is well-established [36, 37, 38, 39], the field of analyzing sentiment through visual cues from non-verbal expres-

sions is still in its early stages. Some studies investigate multimodal sentiment analysis in vlogs, video recordings, and visual behavioral displays [40, 41, 42].

Early research in visual sentiment analysis examined associations between adjectives across images [43], holistic image features [44], and global and local color histograms with SVMs [45]. Alternative mid-level feature representations like Sentribute[46] and SentiBank[47] have also been put forward.

Recent advancements involve the application of Convolutional Neural Networks (CNNs) in visual sentiment analysis [48, 49, 50, 51, 52]. Custom multitask network structures have been proposed [53], along with novel techniques like coupling CNNs with LSTM for sentimentally biased visual captioning [54].

These advancements point to the potential for higher accuracy techniques and increased coverage in visual sentiment analysis, driven by the availability of computer vision models and datasets.

2.1.3 Audio Sentiment Analysis

Audio Sentiment Analysis is a computational process rooted in mathematical methods, designed to identify and categorize the emotional content present in auditory data. This involves extracting pertinent features from audio and mapping them to sentiment labels through a trained model. The procedure encompasses the optimization of model parameters and evaluation using confusion metrics. Ultimately, this systematic approach enables the development of models proficient in accurately interpreting sentiments conveyed through audio recordings.

Speech Analysis for Sentiment: A Growing Focus Within a Longstanding Tradition

While the analysis of speech for emotional and affective cues has a rich history, as evidenced by early work like [55] and recent surveys like [56], the specific area of sentiment analysis within spoken language is a relatively new field.

One of the challenges in this area is the distinction between sentiment and emotion analysis, as highlighted by [57]. While both aim to understand aspects of human expression, sentiment analysis focuses on the overall polarity (positive, negative, or neutral) of an utterance, while emotion analysis delves deeper into specific emotions like anger, joy, or sadness.

Despite this distinction, some research explores the potential of acoustic features in senti-

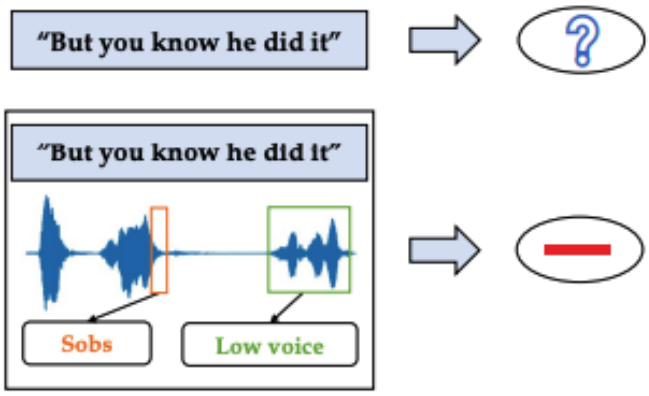

Figure 2.4: An Example of the interplay between the text and audio modalities through cross-modal interaction.[1]

ment analysis. For example, [58] investigate pitch-related features and demonstrate that pitch alone can convey sentiment information.

However, other studies recognize the limitations of relying solely on acoustic cues and explore the role of textual content within speech. Works like [59] and [60] focus on sentiment analysis using the textual content extracted from speech recognition, demonstrating its potential for information retrieval and spoken review analysis.

Furthermore, research by [61] and [62] suggests that sentiment analysis on spontaneous speech data can be effective even with low word recognition rates. This aligns with findings in valence recognition from spontaneous speech [63], suggesting the potential of sentiment analysis even in less-than-perfect speech recognition scenarios.

In essence, while the analysis of emotions and affect in speech has a long history, the specific focus on sentiment within spoken language is a more recent development. Researchers are exploring various approaches, including acoustic features and textual content derived from speech recognition, demonstrating the potential of sentiment analysis in spoken language processing.

2.2 Multimodal Sentiment Analysis

Multimodal Sentiment Analysis (MSA) represents an advanced approach within sentiment analysis that extends its scope beyond textual data to analyze information through multimodal approaches, integrating textual, visual, auditory, and video data." As depicted in Figure 2.5,

the flowchart outlines the various stages involved in Multimodal Sentiment Analysis, starting from data collection to the final application of the analysis results.

The literature on Multimodal Sentiment Analysis can be broadly categorized into two main groups:

(i) Utterance-level models: Algorithms at the utterance level focus on analyzing a target utterance in isolation.

(ii) Inter-utterance contextual models: Contextual algorithms leverage information from neighboring utterances within the entire video, enhancing the understanding of sentiment in a broader context.

This categorization helps in understanding the diverse methodologies and approaches employed in Multimodal Sentiment Analysis, ranging from micro-level analysis to macro-level contextual understanding.

2.2.1 Challenges

The Semantic Gap presents a significant challenge in Multimodal Sentiment Analysis. This gap refers to the difference in representation and interpretation of sentiment across various modalities, such as text, speech, images, and other forms of communication. Bridging this semantic gap is crucial for achieving accurate and reliable sentiment analysis results.

Researchers and practitioners are actively working on addressing this challenge by developing advanced models that incorporate machine learning and natural language processing techniques. Techniques such as transfer learning, deep learning, and multimodal fusion are being employed to enhance the capability of sentiment analysis systems to handle diverse modalities and nuances in sentiment expression effectively.

Several studies have explored sentiment analysis across various contexts, emphasizing the significance of incorporating multiple modalities. Some research has focused on sentiment analysis within movie critiques, while other studies have examined sentiment in YouTube videos, particularly highlighting the importance of visual signals such as facial expressions. Additionally, multimodal strategies have been developed to assess sentiment intensity, demonstrating the effectiveness of combining text, audio, and visual elements. Research in this field also considers facial expressions, speech intonation, and audiovisual emotional analysis, all of which contribute to addressing the challenges of the semantic gap in Multimodal Sentiment

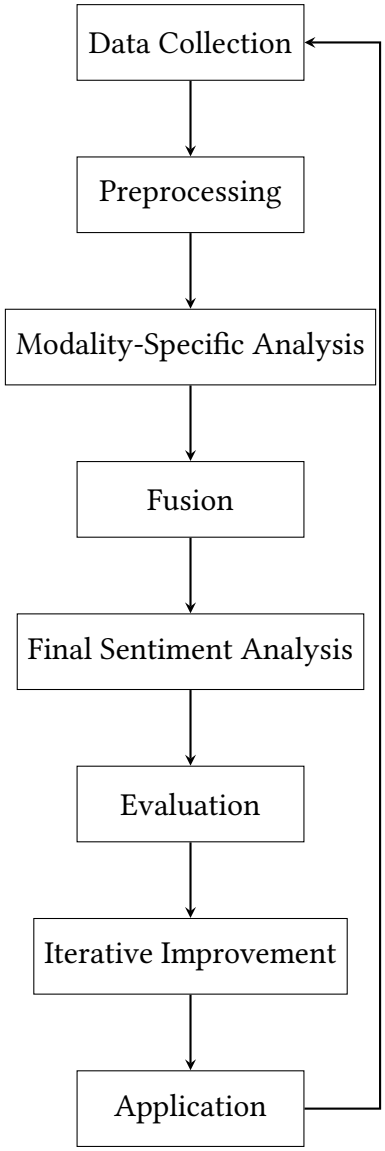

Figure 2.5: Operations involved in Multimodal Sentiment Analysis

Analysis

Integration of Multiple Modalities

Multimodal sentiment analysis diverges from traditional approaches by considering a combination of textual, visual, and auditory information. This can include text-image pairs, text-audio pairs, or text-video combinations.

Deep Learning Architectures

Deep learning models like CNNs [64], RNNs [65], transformers, and their hybrid combinations are commonly utilized to analyze and extract features from diverse modalities. These models are adept at capturing complex relationships and patterns within and between different types of data.

LSTM[66] represents a pivotal advancement in RNN architectures, aimed at mitigating the vanishing gradient problem encountered in traditional RNNs. This problem hinders the effective assimilation of long-range dependencies within sequential data, impacting performance across various domains. LSTM networks excel in processing and forecasting sequential data, exhibiting proficiency across diverse modalities such as time series, textual information, auditory signals, and beyond.

At the core of LSTM networks is the memory cell, a fundamental entity distinct from conventional RNNs. Unlike its predecessors, LSTM cells maintain a memory state capable of retaining information over extended durations, facilitating the assimilation of distant dependencies within input sequences. This capability is crucial for capturing intricate temporal patterns inherent in sequential data.

Integral to the LSTM architecture are its gating mechanisms, comprising forget, input, and output gates. These gates regulate the flow of information within the network, orchestrating the selective retention and discarding of information from previous states, as well as the integration of new input. Leveraging sophisticated activation functions, such as the sigmoid and hyperbolic tangent functions, LSTM networks dynamically modulate the information flow through the gates, thereby governing the evolution of the memory state.

Training LSTM networks involves utilizing Backpropagation Through Time (BPTT)[67], a specialized variant of the backpropagation algorithm tailored for sequential data. Gradient descent methodologies, including Stochastic Gradient Descent (SGD)[68] and its derivatives like Adam[69], are commonly employed to optimize the network parameters, enabling effective

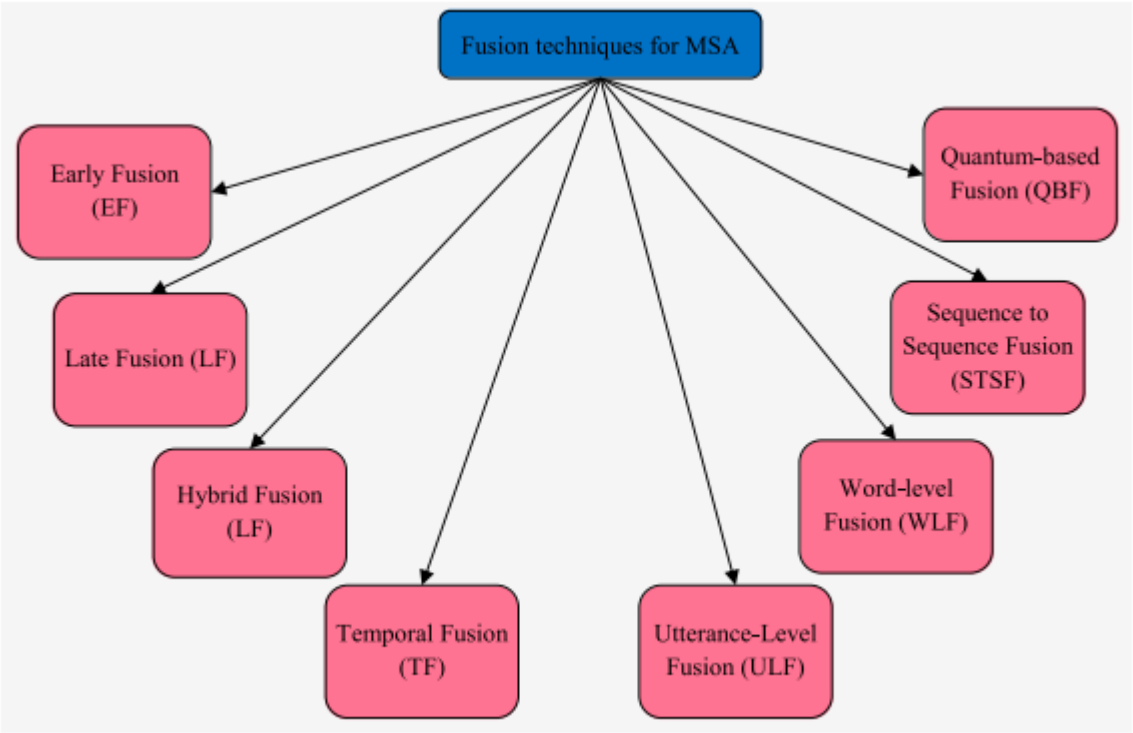

Figure 2.6: Various fusion techniques in Multimodal Sentiment Analysis (MSA) aim to amalgamate a variety of modalities including text, images, and videos[2]

adaptation of LSTM models to diverse tasks.

LSTM networks have gained widespread acclaim for their efficacy across a spectrum of applications. NLP boasts a powerful toolbox, including generating creative text formats, identifying sentiment in written communication, and facilitating communication across languages through machine translation. Furthermore, LSTM networks have demonstrated prowess in time series prediction, speech recognition, image captioning, and health monitoring, underscoring their versatility and utility across disparate domains.

Continual innovation within the LSTM paradigm has led to numerous variants, each introducing refinements to enhance performance and versatility. Gated Recurrent Unit (GRU) and peephole connections represent notable examples of such advancements, augmenting the LSTM architecture with additional mechanisms to bolster its capabilities.

In summary, LSTM networks serve as formidable instruments for modeling sequential data, constituting a cornerstone in contemporary machine learning frameworks. Their resilience to vanishing gradients, coupled with their adeptness in capturing long-range dependencies, renders them indispensable assets across a myriad of cutting-edge applications, solidifying their status as a linchpin in the machine learning landscape.

Fusion techniques in MSA To address the limitations of analyzing sentiment from a single

modality, MSA employs fusion techniques. These techniques integrate information from text, audio, and video, leading to more accurate and robust sentiment analysis. Some common fusion techniques include:

Early Fusion: In this technique, features from different modalities are combined at an early stage of the analysis pipeline, typically before any specific analysis is performed. For example, features extracted from text, audio, and video are concatenated or combined into a single feature vector before being input to a classifier.

Late Fusion: Late fusion involves performing separate analyses on each modality and then combining the results at a later stage. To arrive at a final sentiment prediction, the system can either aggregate confidence scores from different modalities or leverage a meta-classifier to make a combined decision based on the individual classifier outputs.

Feature-level Fusion: In feature-level fusion, extracted features from each modality are combined into a single feature vector before being fed into a classification model. This could involve techniques such as concatenation, weighted summation, or dimensionality reduction techniques like Principal Component Analysis (PCA) or Singular Value Decomposition (SVD).

Decision-level Fusion: Decision-level fusion involves combining the decisions or outputs of individual classifiers trained on each modality. This could involve techniques such as majority voting, weighted voting, or more sophisticated ensemble methods like stacking or boosting.

Hybrid Fusion: Hybrid fusion techniques combine multiple fusion approaches to leverage their respective strengths. For example, a hybrid fusion approach might combine early fusion with decision-level fusion to take advantage of both the feature-level integration and the diversity of multiple classifiers.

Challenges in Multimodal Sentiment Analysis: Coherent integration of information from diverse modalities poses challenges, especially when modalities provide conflicting signals. Handling variations in data types, such as different languages, image resolutions, or audio qualities, requires robust preprocessing techniques.

Datasets for Evaluation: Multimodal sentiment analysis models are often evaluated on benchmark datasets that include diverse sources, such as movie reviews, product reviews, social media posts, and datasets specific to multimodal sentiment like MOSI and MOSEI.

Real-world Applications: Multimodal sentiment analysis finds applications in various

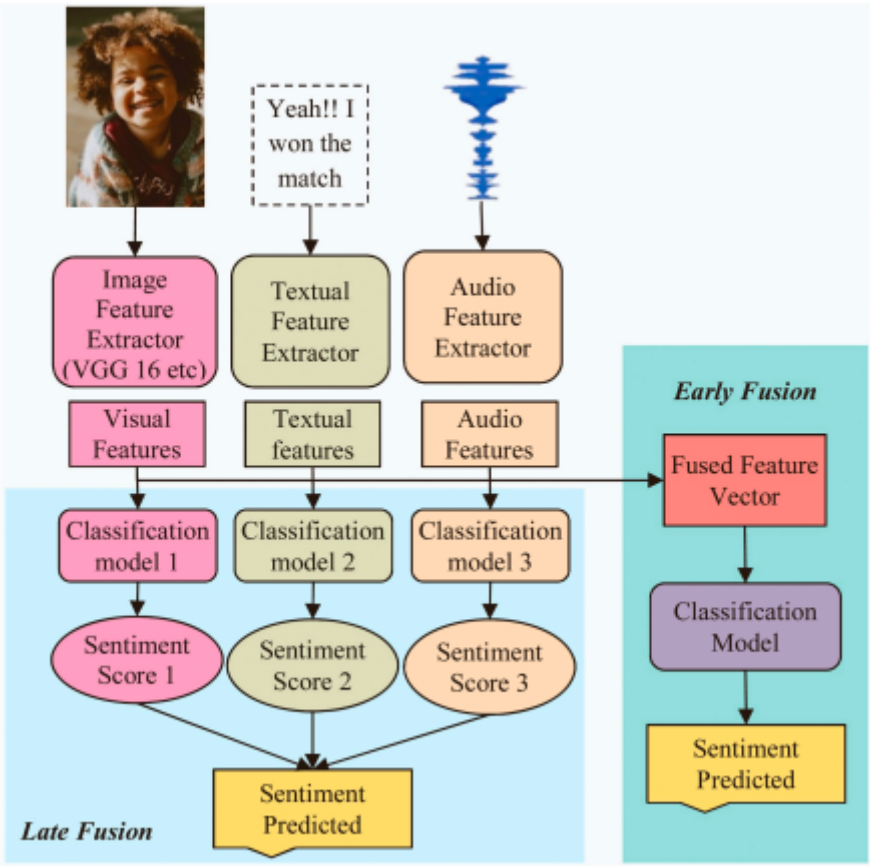

Figure 2.7: Strategy of Early and Late Fusion[2]

domains, including marketing and advertising, where analyzing user responses to multimedia content is crucial. It is also valuable in understanding sentiment expressed in video content, social media posts with images, and audio-based reviews.

2.2.2 Multimodal Invariant and Modality Specific Representations for Multimodal Sentiment Analysis (MISA)

MISA is a multimodal framework designed to learn modality-invariant and modality-specific representations crucial for effective multimodal fusion in affective state prediction[70]. The framework aims to overcome challenges related to modality gaps and the integration of complementary information while minimizing redundancies in multimodal data.

To achieve its objectives, MISA employs a combination of loss functions, including distributional similarity loss for invariant features, orthogonal loss for specific features, reconstruction loss for modality feature representativeness, and task prediction loss. These loss functions work together to learn factorized subspaces for each modality, capturing both shared latent features and modality-specific characteristics.

The framework comprises two primary stages: Modality Representation Learning and Modality Fusion. In the Modality Representation Learning stage, MISA learns invariant and specific features for each modality, providing a comprehensive and disentangled perspective of the multimodal data. These learned representations are then used for fusion in the Modality Fusion stage, where they are integrated to make predictions regarding affective states.

They have also cited pertinent literature from previous studies in their work.

- **Manual Annotation of Opinion Categories in Meetings** [71]: This work extends sentiment analysis beyond textual data by incorporating acoustic and paralinguistic features alongside text-based analysis. It explores how factors like tone, pitch, and other non-verbal cues in speech can contribute to understanding sentiment.
- **Multimodal Subjectivity Analysis of Multiparty Conversation** [72]: This study likely follows a similar vein to [71], investigating deep learning architectures for multimodal sentiment analysis: Fusing text, speech, and nonverbal cues.
- **Extracting Sentiment from Short Speech Reviews** [58]: This research likely extends the exploration of incorporating acoustic and paralinguistic features into sentiment analysis, possibly introducing novel techniques or evaluating the effectiveness of existing methods.
- **Fusion of acoustic and linguistic features for emotion detection** [73]: This work likely explores how acoustic and paralinguistic features can enhance sentiment analysis, potentially focusing on specific applications or domains.
- **Harnessing the Web's Voice: Multimodal Sentiment Analysis** [40]: This study incorporates multimodal cues, including visual cues, for sentiment analysis in product and movie reviews. It directly combines different modalities in an early fusion representation without investigating the relationships between them.
- **Multimodal sentiment analysis of Spanish online videos** [74]: This work also employs multimodal cues for sentiment analysis, particularly in product and movie reviews. However, it also likely focuses on speaker-dependent experiments and lacks analysis of sentiment intensity.

- **Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis** [75]: This research utilizes CNNs for multimodal sentiment analysis. However, unlike the earlier studies, it may focus on utterances rather than opinion segments and could emphasize sentiment polarity rather than intensity. Additionally, it likely employs speaker-dependent experiments.

In summary, MISA offers a straightforward yet adaptable framework that underscores the significance of multimodal representation learning as a preliminary step to multimodal fusion, thereby enhancing the accuracy of affective state predictions in multimodal datasets.

2.2.3 Context-Dependent Sentiment Analysis (Attention Based LSTM)

Context-dependent sentiment analysis [76] introduces a sophisticated Long Short-Term Memory Networks (LSTM)-based model designed to elevate the classification process. This model facilitates the incorporation of contextual information from the surrounding video into utterances, thereby enhancing the nuanced understanding of sentiment within specific contexts. The proposed method surpasses the current state-of-the-art by achieving a remarkable performance improvement of 5-10 percentage points. This accomplishment highlights its ability to effectively capture and interpret intricate contextual subtleties within the data

Beyond its improved performance metrics, the model exhibits a commendable degree of robustness in terms of generalizability. This characteristic is particularly noteworthy as it implies that the model's enhanced sentiment analysis capabilities are not confined to specific scenarios but extend across diverse contexts. The model's strong generalizability signifies its potential for successful transfer to sentiment analysis tasks in new domains and applications, owing to its adaptability.

By effectively leveraging the capabilities of LSTM-based architectures, the paper contributes to advancing the field of sentiment analysis, especially in scenarios where contextual dependencies play a crucial role. The emphasis on contextual information in utterances, coupled with the demonstrated performance gains and generalizability, positions the proposed model as a valuable asset in the realm of context-dependent sentiment analysis, addressing the complexities inherent in understanding sentiment within dynamic and varied environments.

Method: Focus on video sentiment analysis: This is an important direction in the field, as

videos offer richer information than text alone. Analyzing emotions through facial expressions, tone of voice, and body language can provide deeper insights.

Utterance[77] level tagging: Tagging individual speech units with sentiment labels allows for a more granular understanding of the sentiment dynamics within a video. This could be helpful for applications like summarizing user reviews or generating emotional highlights. Recognizing relationships between utterances: This is a key challenge that existing approaches often overlook. Understanding how sentiment shifts and evolves across utterances can provide a more nuanced picture of the overall sentiment.

2.2.4 Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis

The surge in social media and smartphone usage has transformed video into a prominent medium for expressing opinions and sharing product reviews. Platforms like YouTube and Facebook host vast amounts of user-generated content, emphasizing the need for effective emotion recognition and sentiment analysis to comprehend the expressed opinions.

While traditional sentiment analysis primarily focuses on text, videos provide richer information, including facial expressions, tone of voice, and body language. These nonverbal cues offer valuable insights, sometimes even contradicting spoken words.

With this Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis[78], which is a novel approach to utterance-level sentiment analysis in videos, where each spoken segment (utterance) is labeled with its sentiment. The central innovation lies in the proposed attention-based LSTM network (CAT-LSTM), which adeptly models the contextual relationships among utterances within a video.

Unlike existing multimodal sentiment analysis methods that treat utterances in isolation, the CAT-LSTM considers the conversational flow's context. This approach enhances sentiment understanding by capturing subtle shifts in emotion and recognizing the nuanced interplay of sentiments within the video's narrative.

In this, AT-Fusion is a method employing attention mechanisms[24] to fuse information from multiple modalities, and its effectiveness is being assessed by comparing it with a simpler fusion technique using CAT-LSTM for sentiment classification. The comparison may involve

various performance metrics to determine which method yields better results for the given task.

2.2.5 Efficient Low-rank Multimodal Fusion with Modality-Specific Factors (LMF)

The authors [79] of the paper "Efficient Low-rank Multimodal Fusion with Modality-Specific Factors" elaborate on the critical significance of multimodal fusion across various domains such as sentiment analysis, speaker trait analysis, and emotion recognition. They outline the prevalent challenges encountered in multimodal fusion, including the computational complexities and the pressing need for streamlined methodologies capable of efficiently handling multiple modalities.

In their scholarly contribution, the authors propose a novel technique known as Low-rank Multimodal Fusion (LMF) with Modality-Specific Factors. The primary concept driving LMF revolves around harnessing low-rank tensors to fuse multimodal features efficiently, complemented by the incorporation of modality-specific factors designed to capture the unique characteristics inherent in each modality. Through the integration of modality-specific factors, the model can discern distinct representations for different modalities, thus enriching the overall fusion process.

The introduced LMF method seeks to address the limitations associated with existing multimodal fusion techniques by mitigating computational complexity and enhancing performance in tasks necessitating the amalgamation of information from multiple modalities. Leveraging low-rank tensors and modality-specific factors, the authors substantiate the effectiveness of their approach by showcasing competitive results across a diverse range of multimodal tasks.

The proposed LMF method in the paper "Efficient Low-rank Multimodal Fusion with Modality-Specific Factors" introduces a novel approach to multimodal fusion by leveraging low-rank tensors and modality-specific factors. Here are more details on how the LMF method works:

1. **Problem Formulation:** The authors begin by delineating the challenge of multimodal fusion and introduce fusion methods based on tensor representations. While tensors offer expressive power, their scalability to numerous modalities is impeded by computational complexity.

2. **Low-rank Tensor Decomposition:** To address the limitations of tensor-based fusion methods, the LMF method decomposes weights into low-rank factors. This decomposition reduces model parameters and facilitates more efficient computation of tensor-based fusion. By decomposing the weight tensor into low-rank factors, the model leverages parallel decomposition of the input tensor for efficient fusion computation.

3. **Modality-Specific Factors:** Alongside low-rank tensor decomposition, LMF incorporates modality-specific factors to capture unique modality characteristics. By integrating these factors, the model learns distinct representations for modalities, thereby enhancing fusion and task performance.

4. **Efficiency and Performance:** The LMF method undergoes evaluation on tasks such as sentiment analysis, speaker-trait recognition, and emotion recognition. Results demonstrate LMF's competitiveness while significantly reducing computational complexity compared to other tensor-based fusion methods. It performs robustly across low-rank settings and is efficient in training and inference.

5. **Comparison to Other Methods:** LMF surpasses other tensor-based methods by reducing parameters and computational complexity. Leveraging low-rank tensors and modality-specific factors, LMF provides an efficient and effective approach to multimodal fusion in AI tasks.

. They have also cited pertinent literature from previous studies in their work.

- **Facial emotion recognition using multi-modal information** [80] may have delved into techniques or models aimed at recognizing emotions using multimodal information, potentially emphasizing how different modalities contribute to a more precise understanding of emotional states.

- **Multimodal human emotion/expression recognition** [81] might have explored methods for identifying human emotions and expressions, possibly incorporating features from diverse modalities to enhance the recognition process.

- **Extracting sentiment from YouTube movie reviews using audio-visual cues** [82] likely made advancements in the field by investigating sophisticated approaches to emo-

tion recognition, possibly utilizing multimodal data sources and advanced algorithms to enhance the accuracy and robustness of emotion recognition systems.

Collectively, these studies are likely to have made substantial contributions to the realm of emotion recognition, highlighting the significance of multimodal fusion and advanced methodologies in comprehending and interpreting human emotions.

In summary, the LMF method proposed in the paper offers a promising solution to multimodal fusion challenges, combining low-rank tensor decomposition and modality-specific factors to enhance efficiency and task performance across various multimodal scenarios.

2.2.6 Multimodal Fusion with Hierarchical Mutual Information Maximization

The MMIM (Multimodal Fusion with Hierarchical Mutual Information Maximization) framework addresses the complexities encountered in multimodal sentiment analysis tasks and emphasizes the significance of efficiently amalgamating information across diverse modalities, including text, visual, and acoustic inputs. Introduced in the study by Han et al. [3], the framework presents a novel approach to enhancing multimodal fusion for sentiment analysis tasks.

The proposed methodology utilizes Mutual Information (MI) maximization to preserve task-relevant information during the fusion of multimodal inputs. The MMIM framework is built upon two central building blocks: a data fusion module for combining information across modalities, and a mutual information maximization module that optimizes the interdependence between the data and the learned representation. Within the fusion module, a fusion network transforms unimodal representations into a unified fused representation, which is then used for final predictions through a regression multilayer perceptron. The MI maximization module estimates MI lower bounds at both the input and fusion levels to augment the correlation between modalities and enhance prediction accuracy.

By jointly training the model with objectives related to both the primary sentiment analysis task and MI maximization, the MMIM framework proficiently learns the integration of task-relevant information into the fusion outcomes. This approach aims to overcome the limitations of existing methods that primarily focus on back-propagating task loss or manipulate feature

spaces without explicitly considering the preservation of task-specific information during multimodal fusion. The MMIM (Multimodal Fusion with Hierarchical Mutual Information Maximization) framework, as discussed in the MMIM MMSA paper [3], leverages insights from various studies to enhance multimodal fusion in sentiment analysis tasks.

The citation to "Multi-task learning for multimodal emotion recognition and sentiment analysis" [83] likely illustrates the implementation of multi-task learning in multimodal sentiment analysis. The MMIM MMSA paper may have utilized this study to explore how simultaneously learning from multiple related tasks can improve sentiment analysis model performance.

Referencing "Deep variational information bottleneck" [84] emphasizes the concept of information bottleneck in deep learning models. By highlighting the importance of maximizing mutual information during multimodal fusion, the MMIM MMSA paper aims to retain crucial task-related information effectively.

The inclusion of "Speaker identification on the SCOTUS corpus" [85] underscores the significance of audio processing and speaker recognition tasks in multimodal sentiment analysis. This reference highlights the necessity of considering audio modalities in sentiment analysis tasks involving diverse data sources.

By incorporating insights from these referenced papers, the MMIM MMSA paper contributes to advancing the MMIM framework's development and enriches the discourse on effectively fusing multimodal data for sentiment analysis tasks. Overall, the MMIM framework offers a comprehensive solution by hierarchically maximizing Mutual Information, thereby enhancing prediction accuracy and task efficacy in multimodal sentiment analysis.

2.2.7 Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning

In the realm of multimodal sentiment analysis, integrating verbal and nonverbal cues such as text, visual, and acoustic elements is crucial for a comprehensive understanding of sentiment. However, obtaining accurate representations for each modality is challenging due to the lack of independent supervision for individual modalities. The authors address this challenge by introducing a novel approach to learn modality-specific representations through self-supervised multi-task learning .

The authors present a method called Self-MM, designed to obtain separate unimodal super-

vision for each modality in a multimodal sentiment analysis task. By incorporating a unimodal label generation module based on a self-supervised strategy, the model can generate reliable and consistent unimodal supervision without relying on manual annotations. Additionally, they introduce a weight self-adjusting strategy to ensure balanced learning progress across different subtasks during the joint training of multimodal and unimodal tasks.

The proposed Self-MM model comprises one multimodal task and three independent unimodal subtasks. It employs a hard-sharing strategy to share the bottom representation learning network between the multimodal task and different unimodal tasks. For the multimodal task, a conventional architecture for sentiment analysis is used, incorporating a feature representation module, a feature fusion module, and an output module. In the text modality, a pre-trained BERT model is utilized to extract sentence representations. The model also leverages the relative distance value based on the distance between modality representations and class centers to enhance model outputs. Through rigorous experiments on benchmark datasets, the authors demonstrate the effectiveness of their proposed method in acquiring modality-specific representations for multimodal sentiment analysis tasks.

The authors also reference relevant literature to support their work. In 2019, "Multi-modal Machine Learning: A Survey and Taxonomy"[86] provides valuable insights into the field of multimodal machine learning, offering a comprehensive taxonomy and overview of the research landscape. Similarly, in 2020, "CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality" was introduced at ACL 2020. This work introduces a Chinese multimodal sentiment analysis dataset with detailed annotation of modality, significantly contributing to the advancement of research in this area.

2.2.8 A text enhanced transformer fusion network for multimodal sentiment analysis

The introduction of the Text Enhanced Transformer Fusion Network (TETFN) paper [4] emphasizes the critical role of multimodal sentiment analysis, particularly in evaluating sentiments across audio, visual, and textual modalities within video segments. The researchers propose an innovative strategy that prioritizes enhancing the text modality's influence on sentiment analysis by utilizing it as a guiding force for cross-modal mappings. This approach aims to refine the fusion of textual, visual, and acoustic data to improve sentiment analysis

accuracy.

The TETFN model consists of three key components: "Feature Extraction and Contextual Encoders," "Text Enhanced Transformer (TET)," and "Unimodal Label Generation Module (ULGM)." Initially, the model extracts features from each modality and encodes them to capture contextual nuances. It then establishes pairwise cross-modality mappings, with a focus on leveraging text to enrich semantic details within the mappings connecting visual and audio modalities. Finally, a sequence model is employed for sentiment prediction, while unimodal sentiment analysis modules generate labels for individual modalities.

Overall, the TETFN model introduces an innovative paradigm in multimodal sentiment analysis by emphasizing the importance of the text modality and incorporating advanced techniques such as transformer structures alongside pre-trained models like Bert and ViT for feature extraction and contextual encoding. This novel approach aims to address challenges related to integrating diverse modalities for sentiment analysis, ultimately enhancing accuracy and efficacy in sentiment analysis within video content.

They have also cited pertinent literature from previous studies in their work.

- The paper "Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain" [87] introduces a multimodal channel-wise attention transformer model inspired by the brain's multisensory integration mechanisms. The model aims to enhance information fusion from various modalities, similar to how the brain integrates sensory inputs.

- In "Graph-based multimodal fusion with metric learning for multimodal classification" [88], the authors present a graph-based multimodal fusion method that incorporates metric learning for multimodal classification tasks. This method implies a structured approach to integrating information from diverse modalities.

- "Lear ning visual and textual representations for multimodal matching and classification" [89] focuses on learning visual and textual representations for tasks like multimodal matching and classification. It likely explores methods to effectively combine visual and textual information for improved classification accuracy.

- The work "Multi-level multiple attentions for contextual multimodal sentiment analysis" [78] introduces a multi-level multiple attentions approach for contextual multimodal

sentiment analysis, employing attention mechanisms to capture relevant information from different modalities in sentiment analysis contexts.

- Lastly, Building upon the work of Tang et al. (2021) who introduced "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network" [90], this research area explores hierarchical learning methods for sentiment analysis. CTFN utilizes a coupled-translation fusion network to integrate information from various modalities in a hierarchical fashion, aiming to enhance sentiment analysis performance.

These references collectively contribute to the field of multimodal sentiment analysis by exploring diverse fusion techniques, learning representations from different modalities, and enhancing contextual understanding for sentiment analysis tasks. The TETFN model builds upon these works by proposing a novel approach that utilizes text-oriented cross-modal mappings for improved multimodal sentiment analysis.

2.2.9 Context-Dependent Sentiment Analysis (Without Attention Based LSTM)

The paper highlights the overlooked contextual connection among utterances in videos within the current sentiment analysis literature. To bridge this gap, the authors propose an LSTM-based network designed to extract contextual features from utterances in user-generated videos for multimodal sentiment analysis.

The proposed method comprises two primary steps:

1. **Context-Independent Unimodal Utterance-Level Feature Extraction**: Initially, unimodal features are extracted without considering the contextual information of the utterances.

2. **Contextual Unimodal and Multimodal Classification**: Subsequently, the unimodal features are fed into an LSTM network referred to as contextual LSTM. This architecture enables consecutive utterances in a video to share information during the feature extraction process.

By maintaining the sequential order of utterances and facilitating information sharing between consecutive utterances, the proposed framework provides contextual information for the sentiment classification process at the utterance level. This methodology aims to enhance sentiment classification performance compared to traditional frameworks by considering the interdependencies among input utterances in user-generated videos.

The authors introduce an innovative solution to address the contextual relationship among utterances in user-generated videos for sentiment analysis. They propose a Long Short-Term Memory (LSTM) network specifically designed to capture the dependencies and connections among input utterances within a video. This LSTM-based model empowers the utterances to gather contextual information from their surrounding content within the same video, thereby enriching the classification process.

The proposed method begins by independently extracting unimodal features without context. These features are then input into an LSTM network, termed as contextual LSTM, to incorporate contextual information and enable multimodal sentiment analysis. By allowing consecutive utterances in a video to exchange information during feature extraction, the model aims to enhance the performance of utterance-level sentiment classification compared to traditional frameworks.

In summary, the proposed LSTM-based model offers a more holistic and contextually aware approach to sentiment analysis in user-generated videos. It considers the interconnections among utterances and leverages the sequential nature of the data to improve classification accuracy and robustness.

They have also cited pertinent literature from previous studies in their work.

1. Context-Dependent Sentiment Analysis in User-Generated Videos[76]. This study explores context-dependent sentiment analysis in user-generated videos, introducing an LSTM-based model that captures contextual information from surrounding utterances to enhance classification accuracy. The research shows a notable 5-10% improvement over existing methods and emphasizes robustness in terms of generalizability. The work was presented at the 55th Annual Meeting of the Association for Computational Linguistics (ACL) in 2017.

2. Long Short-Term Memory[91]. This influential paper introduces the LSTM neural

network architecture, designed to model long-range dependencies in sequential data. LSTMs have become widely used in various natural language processing tasks, including sentiment analysis, due to their capacity to capture and retain information over extended time intervals.

3. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database[92]. This research introduces the IEMOCAP database, a valuable resource for studying emotional interactions through motion capture data. The database includes dyadic conversations with emotional annotations, providing insights into emotional expressions and dynamics in human communication. It serves as a significant asset for research in emotion recognition and sentiment analysis.

These detailed explanations offer a deeper understanding of the importance and contributions of the referenced papers, providing insights into their respective research areas, methodologies, and implications for context-dependent sentiment analysis in user-generated videos.

2.3 Summary

Chapter 2 provides a comprehensive exploration of sentiment analysis across various modalities, laying the foundation for evaluating eight multimodalities in the thesis. The chapter begins by delving into text sentiment analysis, followed by a detailed examination of image and audio sentiment analysis techniques. Additionally, it explores the concept of multimodal sentiment analysis, highlighting the integration of insights from multiple modalities for a more holistic understanding of emotions.

CHAPTER 3

Methodology

The methodology chapter acts as the roadmap for fulfilling our research objectives. It entails an in-depth examination of the architectures of eight multimodal models applied to the MOSI and MOSEI datasets. Within this chapter, we meticulously delineate the design intricacies of each model, highlighting the integration of textual, auditory, and visual modalities to capture nuanced sentiment expressions. Furthermore, we offer comprehensive insights into the datasets, elaborating on their composition, annotation schemes, and preprocessing steps. By elucidating our methodologies, our goal is to provide readers with a clear comprehension of the research framework, thereby laying the foundation for subsequent exploration of results and discussions.

3.1 Datasets

For the previous dataset mentioned, the YouTube Opinion Dataset [40] consists of 47 YouTube videos annotated for sentiment polarity at the video level by three annotators. This dataset comprises manually transcribed text, automatically extracted audio and visual features, as well as automatically extracted utterances. The MMMO dataset [82] is an extension of the YouTube Opinion Dataset, increasing the number of videos from 47 to 370.

Additionally, the Spanish Multimodal Opinion Dataset [74] is specifically designed for Spanish sentiment analysis, containing 105 videos annotated for sentiment polarity at the utterance level. Utterances are automatically extracted based on extended pauses, with most

videos containing 6-8 utterances. The dataset encompasses a total of 550 utterances. Notably, these datasets lack sentiment intensity annotations, primarily concentrating on polarity. Moreover, they primarily emphasize video or utterance analysis rather than a more granular examination of sentiment, as mentioned earlier.

This lack of sentiment intensity annotations restricts the granularity of sentiment analysis, preventing a more detailed exploration of the emotional nuances present in the data.

Moreover, the dataset places a predominant emphasis on video or utterance-level analysis rather than a more granular examination of sentiment. While video and utterance-level analysis are valuable for understanding overall sentiment trends, the absence of finer sentiment granularity may limit the dataset's applicability in scenarios where a detailed understanding of emotional intensity is essential. Researchers and practitioners seeking to delve into subtle variations in sentiment expression may find the dataset less suited to their specific requirements.

To assess the performance of multimodal systems, it is essential to utilize benchmark datasets in multimodal sentiment analysis (MSA). The MOSI [93] and MOSEI [94] datasets are commonly employed in the field of affective computing and sentiment analysis.

CMU-MOSEI and CMU-MOSI are two pivotal datasets that have made substantial contributions to the field of sentiment analysis and emotion recognition within online videos.

CMU-MOSEI, heralded as the largest dataset for sentence-level sentiment analysis and emotion recognition, features over 65 hours of meticulously annotated video content, encompassing contributions from over 1000 speakers discussing a wide array of 250 topics. Its debut at the 2018 Association for Computational Linguistics marked a pivotal moment in the field, playing a central role in the First Grand Challenge and Workshop on Human Multimodal Language, underscoring its significance and impact on advancing research. The MOSEI dataset comprises a considerable number of videos, enhancing the richness and diversity of the dataset. It proves invaluable for research centered on emotion intensity analysis, facilitating the development of models capable of understanding the nuanced strength of emotional expressions in multimodal content.

- In contrast, CMU-MOSI directs its focus towards opinion-level sentiment intensity within online videos, pioneering sentiment analysis at the utterance level within English videos. MOSI serves as a multimodal dataset encompassing various modalities, including

Table 3.1: Comparison of MOSI and MOSEI Datasets

Aspect	**MOSI**	**MOSEI**
Dataset Size	2199(Number of Clips)	23500+(Number of Clips)
Modalities	Text, Audio, Video	Text, Audio, Video
Sentiment Analysis	Yes	Yes
Emotion Analysis	No	Yes
Intensity Levels	3 (Low, Medium, High)	5 (Very Low, Low, Medium, High, Very High)
Annotations	Human-annotated	Human-annotated
Language	English	English
Applications	Sentiment Analysis, Emotion Recognition	Sentiment Analysis, Emotion Recognition
Temporal Coverage	2007-2015	2017-2021
Data Sources	Online platforms	Various sources
Data Collection	Manual	Manual
Additional Features	-	Audio-Visual Features
Domain	Multimedia	Multimedia
License	Proprietary	Proprietary
Updates	-	Regular updates
Research Use	Extensively used	Widely used

text, audio, and visual (video). Noteworthy for being the largest dataset of its kind upon release, CMU-MOSI comprises 2199 opinion utterances annotated with sentiment, ranging from very negative to very positive and classified into seven Likert steps. This dataset has paved the way for exploring nuanced sentiment expressions, offering researchers a comprehensive resource for comprehending and analyzing sentiment intensity within the realm of online video content.

- Collectively, these datasets serve as invaluable resources in the domain of multimodal sentiment analysis, providing vast and annotated multimedia content for delving into the intricacies of sentiment and emotion across diverse online interactions. The MOSI dataset is enriched with sentiment intensity labels, offering a continuous measure of sentiment strength instead of discrete sentiment classes. This enables a more nuanced analysis of sentiment. Furthermore, the MOSI dataset incorporates annotations for sentiment intensity within each modality individually, enabling researchers to investigate the impacts of text, audio, and visual information on overall sentiment intensity.

3.2 Details of the MOSI dataset

We delve into the MOSI dataset, covering its composition, annotation methodology, and preprocessing. Understanding these details is crucial for our analysis of multimodal sentiment.

3.2.1 Data Source and Composition

Video Clips: The MOSI dataset is compiled from video clips sourced from movie reviews available on YouTube. These clips feature individuals expressing their opinions on diverse topics, creating a rich dataset with varied sentiments.

3.2.2 Modalities

Textual Modality (Transcripts): Transcripts of the spoken content in the video clips are included as the textual modality. These transcripts capture the linguistic aspects of sentiment expression.

Visual Modality (Video Frames): Video frames are provided to represent the visual modality. Each video clip comes with a sequence of frames, allowing the analysis of facial

expressions and non-verbal cues contributing to sentiment.

Audio Modality (Audio Signals): The audio signals extracted from the video clips constitute the audio modality. These signals enable the analysis of vocal intonation and other auditory features contributing to sentiment expression.

3.2.3 Sentiment Annotations

Sentiment Labels: Each video clip within the MOSI dataset is annotated with sentiment labels, offering an indication of the overall sentiment expressed in the video. These labels offer a broad categorization of sentiment.

Sentiment Intensity: Sentiment intensity is represented on a continuous scale. Instead of discrete sentiment classes, this allows for a nuanced understanding of sentiment strength, offering a more detailed perspective on sentiment expression.

3.2.4 Data Partitioning

To prepare the data for model development and evaluation, A common data partitioning strategy involves splitting the dataset into training, validation, and test sets.

Training Set: Utilized for model training.

Validation Set: Employed for hyperparameter tuning during model development.

Test Set: Reserved for assessing model performance on unseen data.

3.2.5 Challenges and Complexity

Diversity: The MOSI dataset presents challenges due to the diversity of speakers, topics, and expressions in the video clips. This diversity enhances the dataset's real-world applicability but also increases the complexity of sentiment analysis tasks.

3.2.6 Metadata

Supplementary Information: Additional metadata, such as speaker details, video duration, and contextual information, may be provided. This supplementary information enhances the contextual understanding of sentiment expression in the dataset.

3.2.7 Availability

The MOSI dataset is commonly released to the public for research purposes, accessible via designated repositories or official channels. This accessibility encourages uniform usage and benchmarking within the realm of sentiment analysis.

3.2.8 Data Preprocessing

Text Preprocessing: This entails cleaning and tokenizing textual transcripts, eliminating stop words, and performing stemming or lemmatization.

Audio Preprocessing: Features, such as spectrograms or Mel-frequency cepstral coefficients (MFCCs), are derived from audio signals.

Visual Preprocessing: Features are extracted from video frames, either through image preprocessing or utilizing pre-trained CNNs.

Modality Alignment: Temporal alignment is ensured to synchronize data from textual, audio, and visual modalities.

Data Integration: Features from diverse modalities are merged into a unified representation, with normalization and standardization applied to ensure equitable contributions from each modality.

Label Processing: Sentiment labels are processed according to the task requirements, potentially converting categorical labels into a numerical format.

Train/Test Split: The dataset is divided into training, validation, and test sets, ensuring there is no data leakage between these partitions.

Optional Data Augmentation: Exploring techniques for augmenting the dataset could enhance its diversity and improve model robustness.

Despite progress, challenges persist in multimodal sentiment analysis, including the semantic gap between modalities, the need for large labeled datasets, and the interpretability of combined features. Researchers continue to address these challenges through innovative model architectures and evaluation methodologies.

3.3 Information about the MOSEI dataset

We delve into the MOSEI dataset, covering its composition, annotation methodology, and preprocessing. Understanding these details is crucial for our analysis of multimodal sentiment.

3.3.1 Data Source and Composition

Video Clips: The MOSEI dataset is curated from video clips sourced from a diverse array of sources, primarily drawn from online platforms hosting movie reviews. These clips showcase individuals expressing opinions on various topics, contributing to a rich dataset characterized by a spectrum of sentiments.

3.3.2 Modalities

Textual Modality (Transcripts): Incorporated as the textual modality are transcripts of spoken content from the video clips. These transcripts meticulously capture the linguistic aspects of sentiment expression.

Visual Modality (Video Frames): In the visual modality representation, each video clip is paired with a sequence of frames. This setup enables the examination of facial expressions and non-verbal cues, which play a significant role in conveying sentiment.

Audio Modality (Audio Signals): Comprising the audio modality are signals derived from the video clips. These audio signals facilitate the examination of vocal intonation and other auditory characteristics that contribute to the overall expression of sentiment.

3.3.3 Sentiment Annotations

Sentiment Labels: Each video clip in the MOSEI dataset undergoes annotation with sentiment labels. These labels serve to categorize the overall sentiment expressed in the video, providing a broad classification.

Sentiment Intensity: Sentiment intensity is represented on a continuous scale, offering a nuanced understanding of sentiment strength. This nuanced representation provides a detailed perspective on sentiment expression, deviating from discrete sentiment classes.

3.3.4 Data Split

The dataset is commonly split into training, validation, and test sets:

- **Training Set:** Used exclusively for training machine learning models.
- **Validation Set:** Employed for fine-tuning hyperparameters during model development.
- **Test Set:** Reserved for evaluating model performance on novel, unseen data.

3.3.5 Challenges and Complexity

Diversity: The MOSEI dataset introduces challenges stemming from the diversity of speakers, topics, and expressions within the video clips. This diversity amplifies the dataset's real-world applicability, albeit at the expense of increased complexity in sentiment analysis tasks.

3.3.6 Metadata

Supplementary Information: Supplementary metadata, such as speaker information, video duration, and contextual details, may also be included. This additional data enriches the thorough comprehension of sentiment expression within the dataset.

3.3.7 Availability

The MOSEI dataset is typically released to the public for research purposes, accessible via designated repositories or official channels. This availability encourages consistent utilization and benchmarking within the realm of sentiment analysis.

3.4 Data Processing

Data annotation stands as a vital initial phase in supervised machine learning. It entails systematically attributing metadata, labels, or tags to individual elements within a dataset. The objective is to render the data understandable, identifiable, and practically beneficial for machine learning algorithms. Particularly in supervised learning, where models are trained on labeled instances to carry out predictions or execute specific tasks, data annotation assumes a pivotal role.

This annotation process is characterized by the dichotomy between manual and automated methodologies. Manual annotation involves human annotators meticulously assigning labels or descriptors to data points, while automated annotation leverages algorithms or pre-existing models for the expeditious assignment of metadata. Quality control mechanisms, including inter-annotator agreement assessments for human annotators and validation checks for automated annotations, are implemented to ensure the fidelity and accuracy of the annotated data.

The applications of data annotation are extensive, covering a range of machine learning domains including image recognition, natural language processing, and speech recognition. Through providing datasets with ground truth labels, data annotation supports the training of supervised learning models, empowering them to recognize patterns and make informed predictions on new, unseen data. Essentially, the methodical annotation of data acts as a fundamental cornerstone in building robust and effective machine learning models.

Dataset	Total Samples	Training Set	Validation Set	Test Set
CMU MOSI	2,199	1,284	229	686
CMU MOSEI	23,500+	16,265	1,869	4,647

Table 3.2: Data Splits for CMU MOSI and CMU MOSEI Datasets

3.5 MOSEI and MOSI Datasets for Research on Multimodal Sentiment Analysis

The MOSEI and MOSI datasets are fundamental resources for investigating multimodal sentiment analysis. Compiled by prominent research institutions, these datasets present a range of multimodal data, annotated with sentiment labels and intensity scores. They offer valuable insights into the intricacies of emotion expression across various modalities, fueling progress in research on multimodal sentiment analysis.

3.5.1 Multimodal Representations

Inherent Multimodality: MOSEI and MOSI exhibit a fundamental multimodal structure, encapsulating textual, visual, and audio modalities. This inherent multimodality affords

researchers a unique vantage point for sentiment analysis, allowing for the concurrent examination of linguistic, facial, and vocal modalities.

3.5.2 Granular Sentiment Annotations

Sentiment Intensity Annotation: Distinguished by sentiment intensity annotations, both datasets facilitate a nuanced comprehension of sentiment strength. This continuous annotation paradigm contrasts with categorical alternatives, offering heightened granularity in sentiment expression analysis.

3.5.3 Scale and Diversity

Robust Dataset Size: MOSEI and MOSI offer a sizable collection of video clips, furnishing a rich and varied dataset for training and assessing sentiment analysis models. The extensive scale and diversity of these datasets substantially enhance the development of models with increased robustness and generalizability.

3.5.4 Real-world Pertinence

Speaker and Topic Diversity: The datasets exhibit a spectrum of speakers expressing opinions on diverse subjects, enhancing their real-world applicability. This diversity underscores the datasets' relevance in addressing sentiment analysis challenges across varying domains and contextual landscapes.

3.5.5 Benchmarking and Comparative Evaluation

Ubiquitous Benchmark Utility: MOSEI and MOSI have ascended to ubiquity as benchmark datasets within sentiment analysis research. Their pervasive utilization facilitates rigorous benchmarking and systematic comparison, ensuring consistency and methodological transparency in research outcomes.

3.5.6 Research Catalysis

Pioneering Research Impact: Both datasets have engendered substantive advancements in the domain of multimodal sentiment analysis. Their availability has catalyzed innovation,

propelling the evolution of novel models and methodologies within the field.

3.5.7 Accessibility and Community Integration

Public Accessibility: MOSEI and MOSI are publicly accessible, fostering collaborative research endeavors and communal engagement. This accessibility not only supports widespread utilization but also ensures a collective benchmarking standard within the sentiment analysis research community.

3.5.8 Continuous Iterative Development

Iterative Enhancement: The datasets are subject to continuous iterative refinement, signifying an ongoing commitment to their currency and pertinence in contemporary sentiment analysis research.

In conclusion, the preference for MOSEI and MOSI as datasets is based on their inherent multimodal composition, detailed sentiment annotations, extensive dataset scale, and significant impact on advancing sentiment analysis research."

3.6 Deep Learning Multimodals Framework

After a comprehensive review of the prior literature, we have chosen eight baseline deep learning models for evaluation. The detailed descriptions of these baseline deep learning models are presented in the following sections.

3.6.1 MISA

The functioning of MISA can be categorized into two main stages: Modality Representation Learning and Modality Fusion. The entire framework is illustrated in Figure 3.1.

For the MISA model, a learning rate of 1×10^{-4}, batch size of 32, and the Adam optimizer were used for Modality Representation Learning. For Modality Fusion, a lower learning rate of 5×10^{-5}, batch size of 64, and the AdamW optimizer were employed.

For generating **Utterance-level representations**, each modality $m \in l, v, a$ has its utterance sequence, represented as U$m \in$ R$Tm \times dm$, mapped to a fixed-sized vector u$m \in$ Rdh. This

process is accomplished using a stacked bi-directional LSTM. The final hidden representations from the LSTM, combined with a fully connected dense layer, yield the vector $\mathbf{u}_m$:

$$\mathbf{u}_m = \text{sLSTM}\left(\text{U}_m; \theta_m^{lstm}\right) \tag{3.1}$$

In the context of Modality-Invariant and -Specific Representations, each utterance vector $\mathbf{u}_m$ undergoes projection into two distinct representations. The first, termed the modality-invariant component, is devised to learn a shared representation within a common subspace, guided by distributional similarity constraints [95]. This constraint is strategically employed to minimize the heterogeneity gap, deemed a desirable attribute for multimodal fusion. The second representation is the modality-specific component, aimed at capturing the unique characteristics inherent to each modality. Throughout the paper, it is argued that the combination of both modality-invariant and -specific representations is crucial for achieving effective fusion, providing a comprehensive perspective. The primary objective of this study is to facilitate the learning of these representations.

When presented with the utterance vector $\mathbf{u}_m$ for modality m, the model learns the hidden modality-invariant $\left(\mathbf{h}_{c_m} \in \mathbb{R}^{d_h}\right)$ and modality-specific $\left(\mathbf{h}_{p_m} \in \mathbb{R}^{d_h}\right)$ representations through the encoding functions:

$$\mathbf{h}_m^c = E_c\left(\mathbf{u}_m; \theta^c\right), \quad \mathbf{h}_m^p = E_p\left(\mathbf{u}_m; \theta_m^p\right) \tag{3.2}$$

In their research, they introduced a multimodal affective framework named MISA, which dissects modalities into modality-invariant and modality-specific features, followed by fusion for predicting affective states. Despite employing simple feed-forward layers, MISA proves highly effective, showcasing substantial advancements over existing approaches in tasks like multimodal sentiment analysis and humor detection. Their exploratory analysis identifies positive traits, such as reducing the modality gap, accomplished by the representation learning functions, thereby obviating the necessity for intricate fusion mechanisms. The researchers emphasize the significance of prioritizing representation learning before fusion and validate its efficacy through rigorous experimentation.

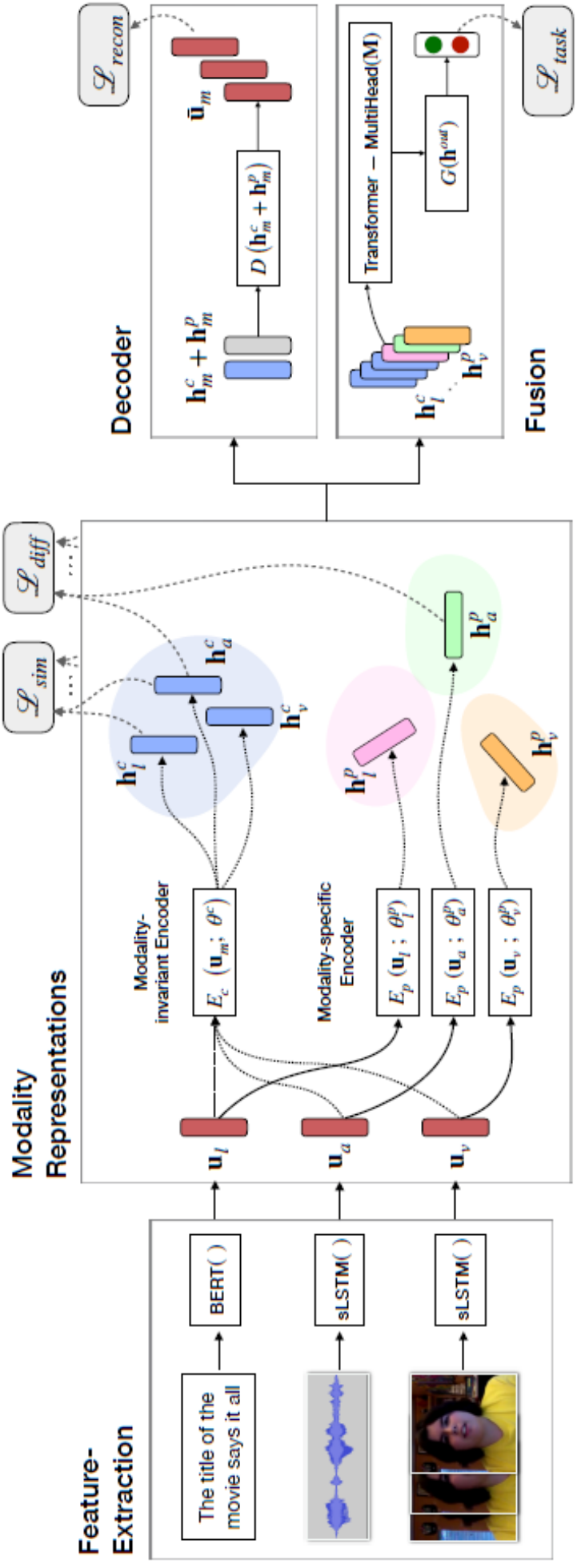

Figure 3.1: In MISA, utterance representations are split into modality-agnostic and modality-specific subspaces to enhance reconstruction and prediction accuracy.

3.6.2 MMIM

For the MultiModal InfoMax (MMIM) model, a learning rate of 2×10^{-4}, batch size of 16, and the Adam optimizer were used for Modality Representation Learning. For Modality Fusion, a learning rate of 1×10^{-4}, batch size of 32, and the AdamW optimizer were employed.

The architecture of the MultiModal InfoMax (MMIM) model, as depicted in Figure 3.2, consists of several key components designed to optimize multimodal fusion for sentiment analysis. Here is a breakdown of the architecture:

Input Processing:

The raw input data from various modalities (text, visual, and acoustic) undergoes initial processing by a feature extractor for visual and acoustic data, and a tokenizer for text data. This process transforms the raw input into numerical sequential vectors.

Modality Encoding:

The model encodes the sequential input from each modality (text, visual, and acoustic) into unit-length representations denoted as h_m, where m represents the modality (text, visual, or acoustic).

Fusion and MI Maximization:

The architecture consists of two main parts: Fusion and Mutual Information (MI) Maximization.

Fusion Part:

A fusion network F, consisting of stacked linear-activation layers, converts the unimodal representations into a fusion result represented as Z.

Subsequently, the fusion result Z is fed into a regression multilayer perceptron (MLP) to make final predictions for the primary sentiment analysis task.

MI Maximization Part:

This section emphasizes maximizing Mutual Information between pairs of unimodal inputs and between the results of multimodal fusion and unimodal inputs.

The MI lower bounds are estimated at two levels—input level and fusion level—and are enhanced to improve the fusion process and prediction accuracy.

Task and MI-Related Losses:

The Fusion and MI Maximization components operate simultaneously to produce task-related losses and MI-related losses for back-propagation during training..

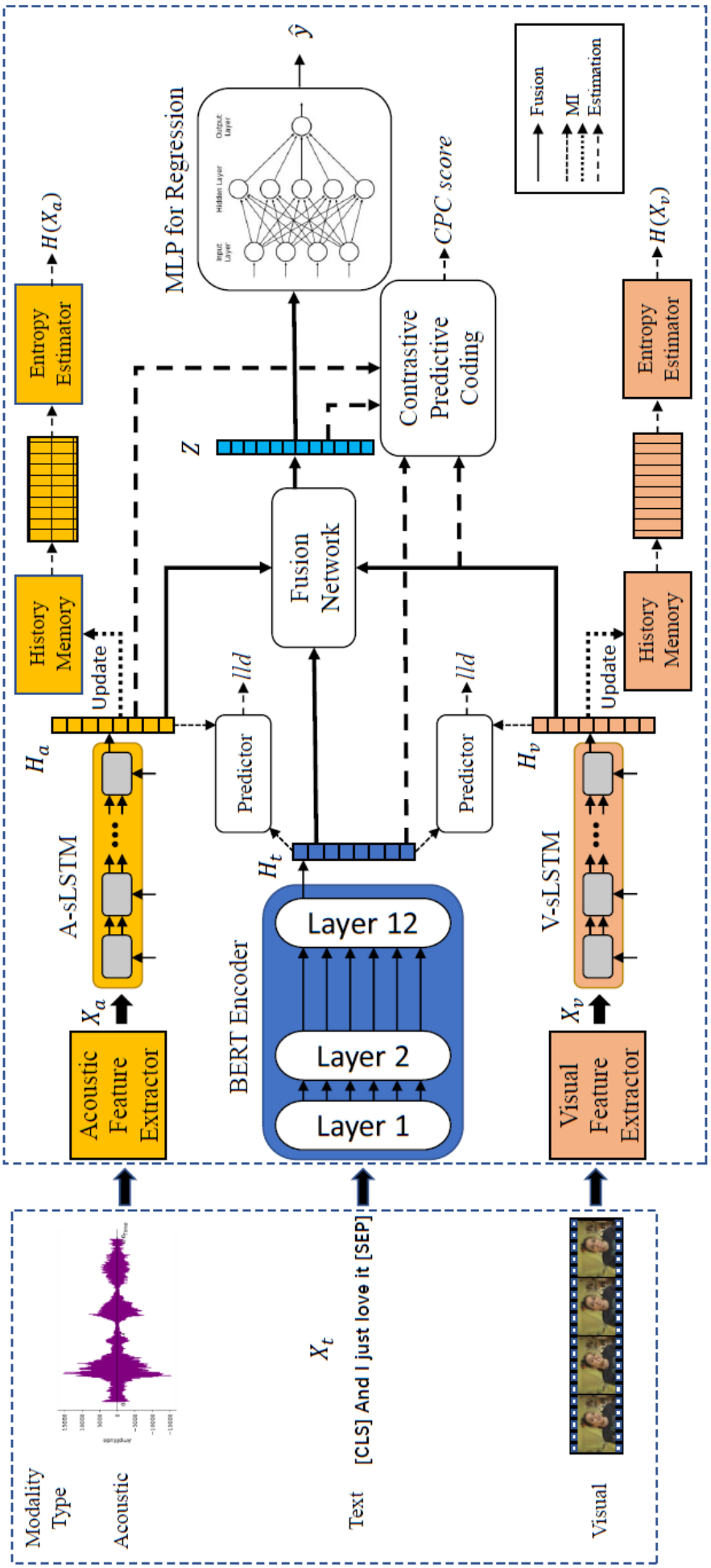

Figure 3.2: The structure and operation of the MMIM model [3] are detailed through its interconnected components and design.

By infusing task-related information into fusion results and improving prediction accuracy, the model learns to optimize multimodal fusion for sentiment analysis effectively.

The study presents a model that hierarchically maximizes mutual information (MI) within a multimodal fusion pipeline. This model integrates two MI lower bounds, one for unimodal inputs and another for the fusion stage, respectively..

1. For unimodal inputs X and Y, the lower bound for MI is given by:

$$I(X;Y) \geq E_{p(x,y)}\left[\log \frac{q(x,y)}{q(x)q(y)}\right]$$

   where $q(x,y)$ is an estimation of the joint distribution of X and Y, and $q(x)$ and $q(y)$ are estimations of the marginal distributions.

2. For the fusion stage involving multimodal data, the MI lower bound is expressed as:

$$I(Z;X,Y) \geq E_{p(z,x,y)}\left[\log \frac{q(z|x,y)}{q(z)}\right]$$

   where Z represents the fused representation, and $q(z|x,y)$ and $q(z)$ are estimations of the conditional and marginal distributions of Z, respectively.

To address the computational complexity and intractability of these lower bounds, the study devises precise, efficient, and robust estimation methods. These techniques ensure uninterrupted training and lead to improved test outcomes.

Subsequently, extensive experiments are conducted on two datasets, followed by an ablation study. The results from these experiments confirm the effectiveness of the proposed model and underscore the importance of the MI maximization framework.

Moreover, the researchers provide visualizations of the losses and present representative examples to provide a more comprehensive insight into the workings of the model.

$$L_{\text{main}} = L_{\text{task}} + \alpha L_{\text{MI}}$$

where L_{main} is the main loss comprising task loss L_{task} and MI-related loss L_{MI}, and α is a hyperparameter controlling the impact of MI maximization.

3.6.3 LSTM W/O Attn

For the LSTM without Attention (LSTM W/O Attn) model, a learning rate of 3×10^{-4}, batch size of 64, and the RMSprop optimizer were used for training. No separate modality fusion process was applied in this model.

The approach outlined for analyzing sentiment in user-generated videos within specific contexts utilizes LSTM-based models. These models are designed to capture the contextual nuances from surrounding speech segments without incorporating attention mechanisms.

- Define $X = \{x_1, x_2, ..., x_M\}$ as the features extracted from individual speech segments without considering contextual factors.
- These features undergo processing through LSTM layers to capture temporal relationships within each segment.
- The equations that govern the LSTM cells are as follows:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

  In this context, f_t, i_t, and o_t represent the forget, input, and output gates, respectively. c_t denotes the cell state, h_t is the hidden state, x_t is the input at time step t, and W and b refer to the weight matrices and bias terms.

- The context-independent unimodal features $Z = \{z_1, z_2, ..., z_M\}$ are inputted into the LSTM network to capture contextual cues between speech segments.
- This LSTM structure enables consecutive speech segments in a video to exchange information during feature extraction.
- The output h_i from the LSTM at time step i is used for classification via a softmax layer.

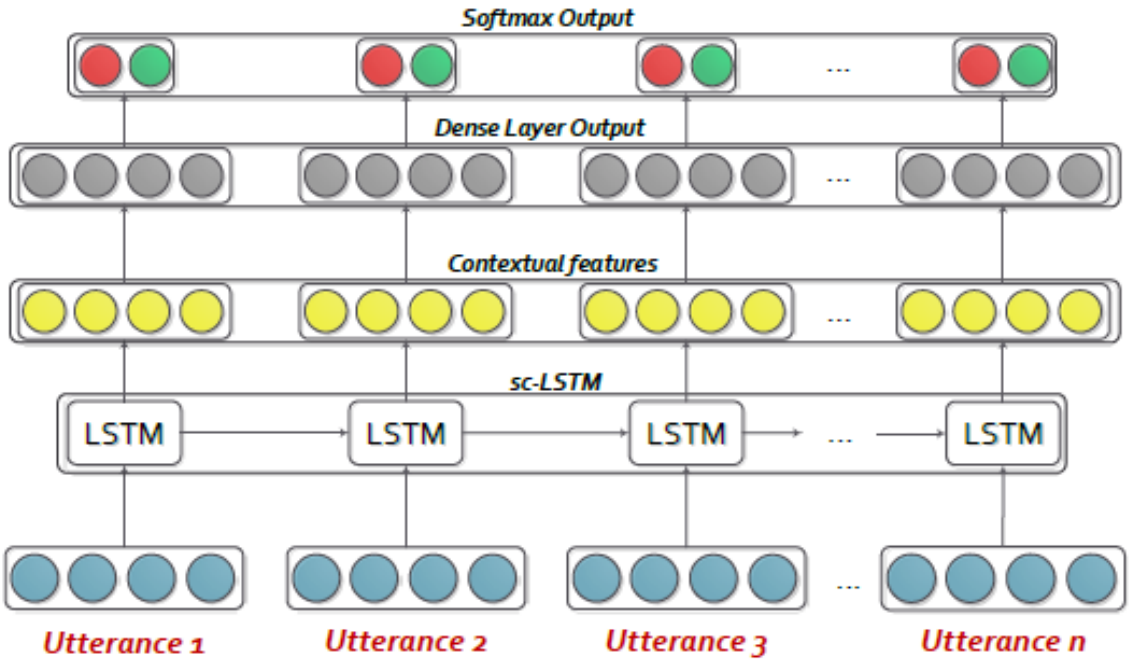

Figure 3.3: The Contextual LSTM network (Figure 3) utilizes a unidirectional LSTM for feature processing, followed by a dense layer for transformation and a softmax layer for final output.

- Training involves optimizing the LSTM network parameters to minimize a sentiment classification loss function.
- The LSTM network is trained to discern contextual dependencies among speech segments, thereby enhancing sentiment analysis performance.

By employing LSTM architectures and considering inter-segment dependencies without incorporating attention mechanisms, this framework improves sentiment analysis accuracy in user-generated videos. The included mathematical formulations provide an in-depth understanding of the context-aware sentiment analysis architecture.

Figure 3.3 depicts the concept of context-independent unimodal feature extraction.

3.6.4 LMF

While tensors are renowned for their expressive power, they often face difficulties in efficiently managing a large number of modalities. To overcome this issue, the proposed model decomposes weights into low-rank factors, significantly reducing the model's parameter count. This decomposition process is expedited by utilizing parallel decomposition of low-rank weight tensors and input tensors for tensor-based fusion, achieving linear scalability concerning the number of modalities.

For the LMF (Low-rank Multimodal Fusion) model, a learning rate of 1×10^{-3}, batch size of 32, and the Adam optimizer were used for both Modality Representation Learning and Modality Fusion.

The fusion process involves computing a low-rank tensor representation h using the unimodal representations z_a, z_v, z_l and the modality-specific factors $w_m^{(i)}$. Mathematically, the fusion operation is expressed as:

$$h = \sum_{i=1}^{r} \left(\prod_{m=1}^{M} (w_m^{(i)})^T z_m \right)$$

where:

- h signifies the fused multimodal representation.
- r represents the rank of the low-rank tensor.

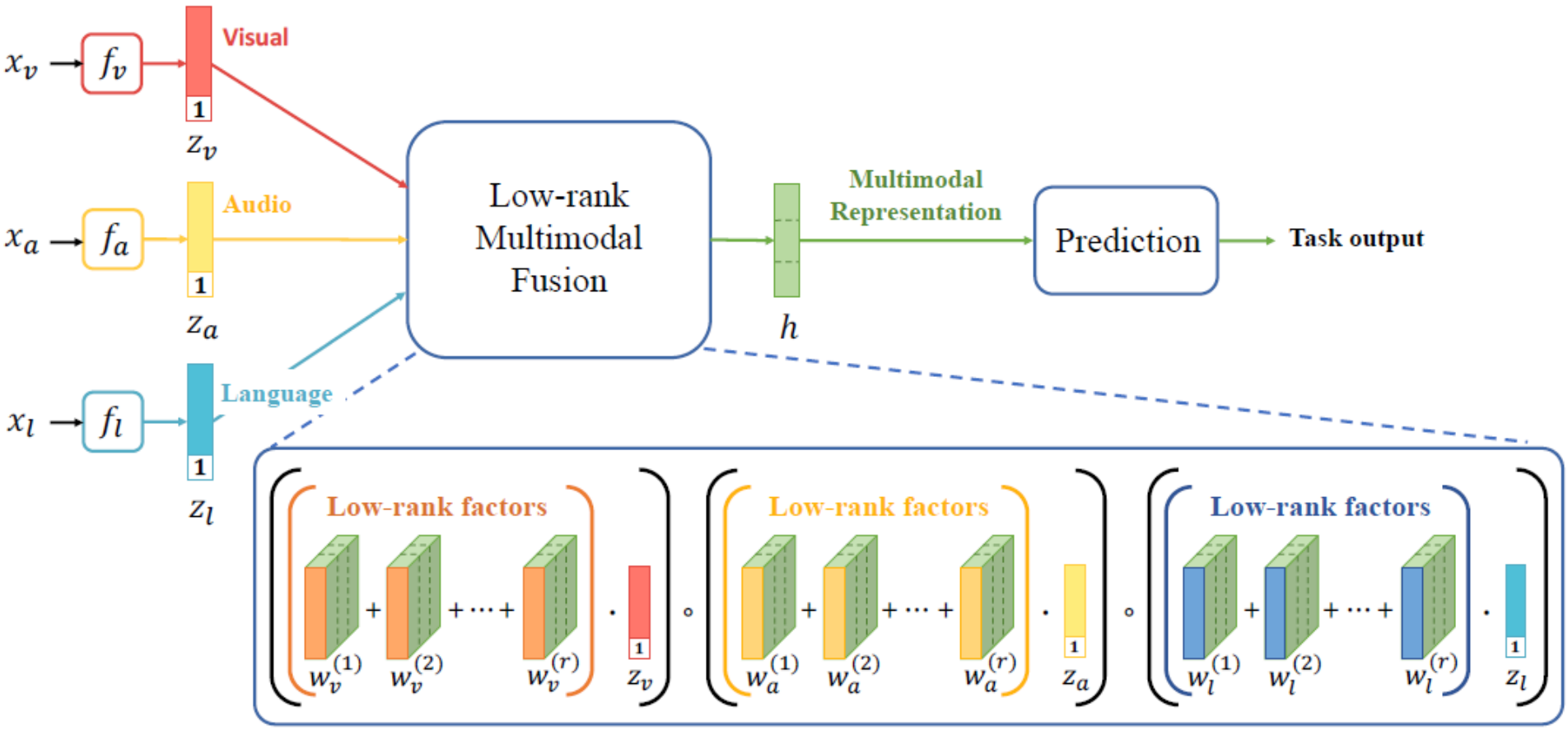

Figure 3.4: The proposed Low-rank Multimodal Fusion (LMF) model (inspired by Liu et al., 2018 [79]) extracts modality-specific features through separate sub-networks in the first stage. These features are then fused using low-rank multimodal fusion in the second stage to create a comprehensive representation for prediction tasks.

- M represents the number of modalities.
- $w_m^{(i)}$ represents the modality-specific factor for modality m and rank i.
- z_m denotes the unimodal representation for modality m.

Efficiency and Differentiability

The fusion operation is fully differentiable, enabling the learning of parameters $w_m^{(i)}$ through end-to-end back-propagation. This differentiability ensures that the model can be trained efficiently using gradient-based optimization methods. Additionally, the modular approach in the fusion operation allows straightforward scalability to any number of modalities by introducing additional modality-specific factors. Moreover, the fusion operation boasts computational efficiency by avoiding the explicit computation of large input tensors and linear transformations.

Figure 3.4 illustrates the architecture of the low-rank multimodal fusion process.

3.6.5 SELF MM

The architecture of the Self-MM model includes a multimodal task along with three independent unimodal subtasks. It employs a hard-sharing approach, utilizing a shared bottom representation learning network for both the multimodal and unimodal tasks. Now, let's explore the architectural details through mathematical expressions.

For the SELF MM model, a learning rate of 5×10^{-4}, batch size of 16, and the Adam optimizer were used for Modality Representation Learning. For Modality Fusion, a learning rate of 3×10^{-4}, batch size of 32, and the AdamW optimizer were employed.

Multimodal Task

The multimodal task involves the following components:

- **Feature Representation Module:** The text modality leverages a pre-trained 12-layer BERT model to extract sentence representations. We can denote the sentence representation derived from BERT as F_t.
- **Feature Fusion Module:** The feature fusion module merges the representations from various modalities. We can represent the resulting multimodal representation as F_m.

- **Output Module:** The output module analyzes the fused representation to generate predictions for the sentiment analysis task.

Unimodal Subtasks

Each unimodal subtask focuses on a specific modality (text, audio, vision) and aims to learn modality-specific representations independently.

Let's denote the representations learned from the unimodal tasks as F_{tu} for text, F_{au} for audio, and F_{vu} for vision.

Hard-Sharing Strategy

The foundational representation learning network is shared between both the multimodal task and the unimodal tasks. This shared structure guarantees that the model can efficiently learn common features across diverse modalities.

Mathematical Expressions

The multimodal representation F_m is obtained by fusing the modality-specific representations: $F_m = \text{Fuse}(F_t, F_a, F_v)$

The unimodal representations F_{tu}, F_{au}, and F_{vu} are learned independently through their respective tasks.

The model aims to optimize a joint objective function that balances the learning progress across all tasks:

$$\mathscr{L}_{\text{joint}} = \lambda_1 \mathscr{L}_{\text{multimodal}} + \lambda_2 \mathscr{L}_{\text{text}} + \lambda_3 \mathscr{L}_{\text{audio}} + \lambda_4 \mathscr{L}_{\text{vision}}$$

where $\mathscr{L}_{\text{multimodal}}$ is the loss for the multimodal task, and $\mathscr{L}_{\text{text}}$, $\mathscr{L}_{\text{audio}}$, and $\mathscr{L}_{\text{vision}}$ are the losses for the text, audio, and vision tasks, respectively. The hyperparameters λ_1, λ_2, λ_3, λ_4 control the importance of each task in the joint optimization.

For a visual representation of the architecture, please refer to Figure 3.5.

To thoroughly examine the integration of attention mechanisms into the hierarchical LSTM architecture for sentiment analysis in user-generated videos, let's delve into the mathematical details and consider the following components:

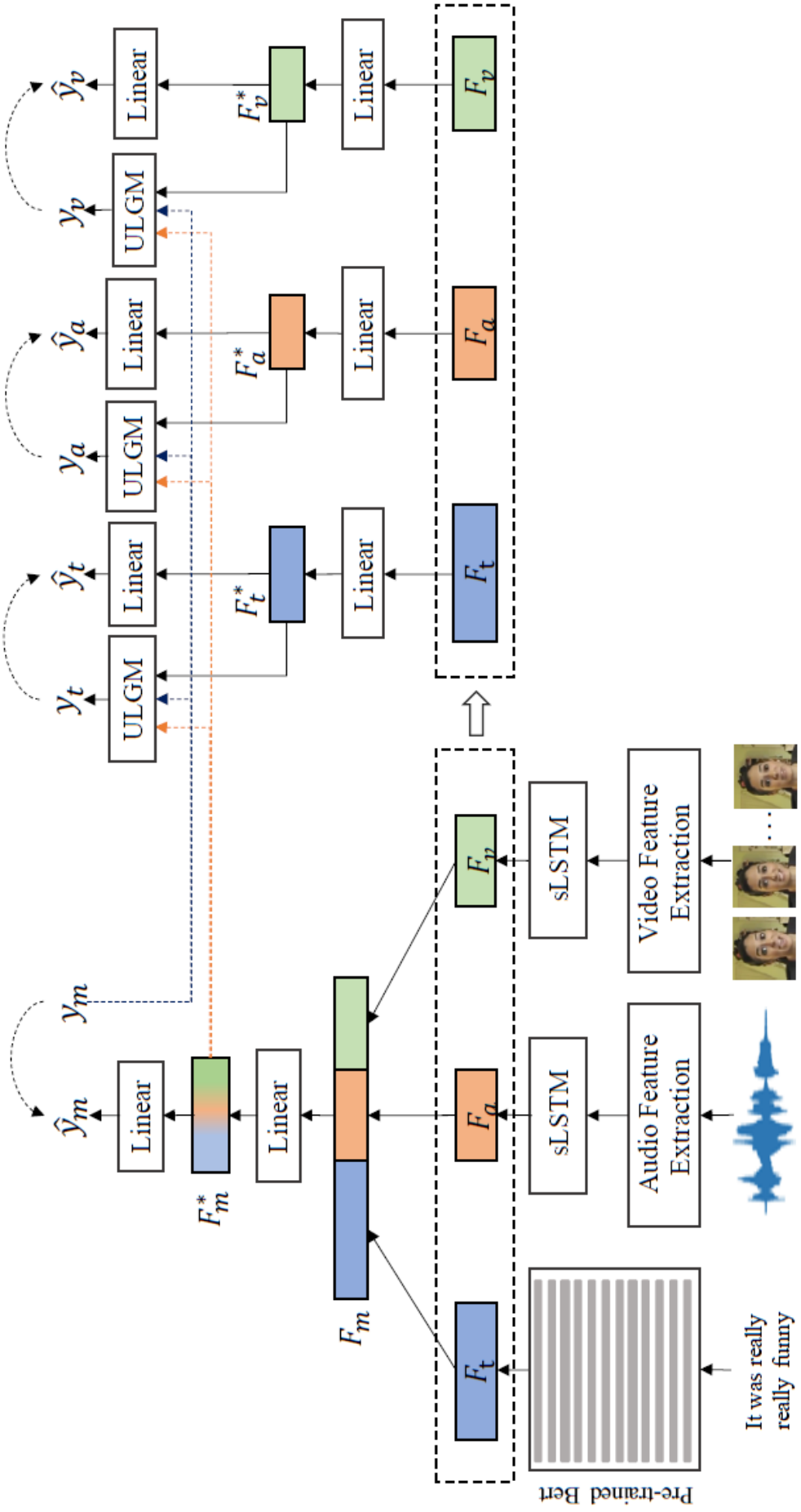

Figure 3.5: Self-MM employs a structure (Figure X) for both multimodal sentiment prediction ($\hat{y}^m$) and unimodal tasks (text, audio, video). Human annotations guide the multimodal task, while self-supervised learning provides supervision for unimodal tasks (y_t, y_a, y_v).

1. **Attention Mechanism Formulation**:

   - During each time step t of the LSTM network, the attention mechanism calculates attention weights $\alpha_{i,t}$ for every utterance i within the video sequence. These attention weights are determined by evaluating the similarity score between the current representation of an utterance and a context vector [T7]..
   - The attention weight $\alpha_{i,t}$ can be computed using a softmax function over the relevance scores:

     $$\alpha_{i,t} = \frac{\exp(e_{i,t})}{\sum_{j=1}^{L} \exp(e_{j,t})}$$

     where $e_{i,t}$ is the relevance score for utterance i at time step t, and L is the total number of utterances in the video.

2. **Context Vector Calculation**:

   - The context vector c_t is calculated by taking a weighted sum of the utterance representations, where the weights are determined by the attention mechanism:

     $$c_t = \sum_{i=1}^{L} \alpha_{i,t} \cdot x_{i,t}$$

     where $x_{i,t}$ is the feature vector representation of utterance i at time step t.

3. **Integration with LSTM**:

   - The context vector c_t is then integrated into the LSTM cell operations to modulate the input or hidden state. This integration allows the LSTM network to focus on relevant utterances based on the attention weights:

     $$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_t + b_i)$$
     $$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_t + b_f)$$
     $$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$
     $$g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + W_{cg}c_t + b_g)$$
     $$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
     $$h_t = o_t \odot \tanh(c_t)$$

where c_t is the context vector computed using attention weights, and x_t, h_{t-1} are the input and hidden state of the LSTM cell at time step t.

4. **Training with Attention**:
    - During training, the attention mechanism parameters are jointly optimized with the LSTM network parameters to learn the optimal attention weights that focus on relevant utterances for sentiment analysis [T12].
    - The model trains to allocate greater attention weights to informative utterances, thereby enhancing the overall performance of the architecture in capturing subtle sentiment nuances in user-generated videos.

By incorporating attention mechanisms into the hierarchical LSTM architecture as described above, the model dynamically focuses on important utterances and enhances its understanding of contextual information within the video sequence, leading to improved sentiment analysis results.

3.6.6 TETFN

Exploring the Architecture of Textual-Enhanced Transformer Fusion Network. For the TETFN model, a learning rate of 2×10^{-4}, batch size of 32, and the Adam optimizer were used for Modality Representation Learning. For Modality Fusion, a learning rate of 1×10^{-4}, batch size of 64, and the AdamW optimizer were employed.

- **Feature Extraction and Contextual Encoders:**
    - Let's denote the extracted features from text, visual, and acoustic modalities as $X_{\text{text}} \in \mathbb{R}^{T_t \times d_t}$, $X_{\text{visual}} \in \mathbb{R}^{T_v \times d_v}$, and $X_{\text{acoustic}} \in \mathbb{R}^{T_a \times d_a}$, respectively, where T_t, T_v, and T_a are the sequence lengths, and d_t, d_v, and d_a are the feature dimensions.
    - The feature extraction and contextual encoders process these features to capture contextual information within each modality, resulting in encoded representations denoted as $H_{\text{text}} \in \mathbb{R}^{T_t \times d_h}$, $H_{\text{visual}} \in \mathbb{R}^{T_v \times d_h}$, and $H_{\text{acoustic}} \in \mathbb{R}^{T_a \times d_h}$, where d_h is the dimensionality of the encoded representations.
- **Text Enhanced Transformer (TET):**

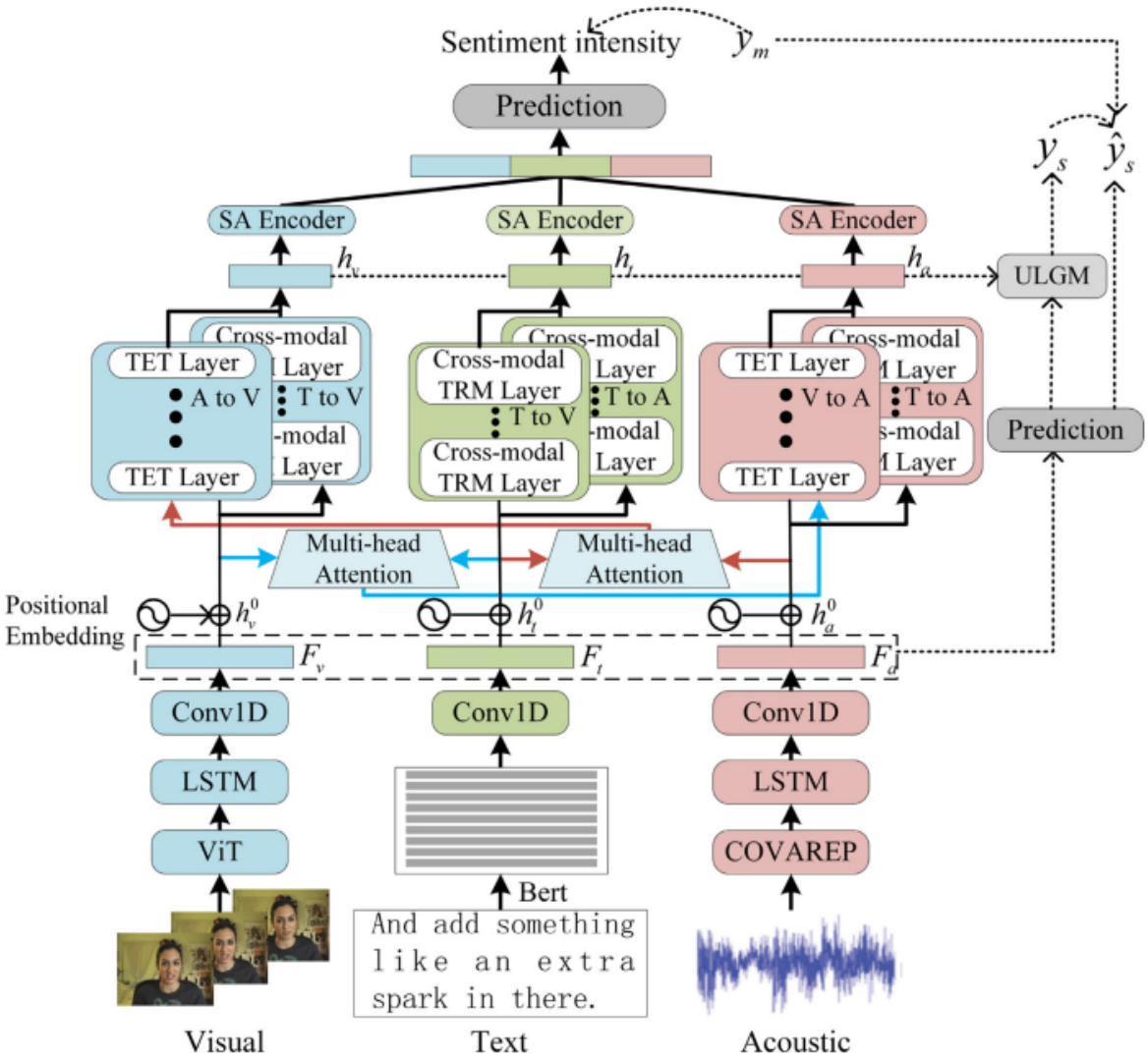

Figure 3.6: The core framework of TETFN [4] consists of three-module architecture for sentiment analysis: feature extraction/encoders, the TET, and the ULGM.

- The TET module incorporates textual information into audio and visual modalities using a multi-head attention mechanism. Let's represent the query, key, and value matrices for each modality as Q_{text}, K_{text}, V_{text}, Q_{visual}, K_{visual}, V_{visual}, Q_{acoustic}, K_{acoustic}, and V_{acoustic}. The multi-head attention mechanism computes attention scores between the query and key matrices and generates weighted combinations of the values.
- The output of the attention mechanism can be expressed as:

$$\text{output}_i = \text{Multihead}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

where output_i represents the output of the i-th head, and d_k is the dimensionality of the keys and queries.

- **Unimodal Label Generation Module (ULGM):**
  - After processing through the TET module, each modality obtains enhanced representations denoted as $H'_{\text{text}} \in \mathbb{R}^{T_t \times d_h}$, $H'_{\text{visual}} \in \mathbb{R}^{T_v \times d_h}$, and $H'_{\text{acoustic}} \in \mathbb{R}^{T_a \times d_h}$. The ULGM generates unimodal labels using these enhanced representations, denoted as $Y_{\text{text}} \in \mathbb{R}^{T_t}$, $Y_{\text{visual}} \in \mathbb{R}^{T_v}$, and $Y_{\text{acoustic}} \in \mathbb{R}^{T_a}$, capturing differentiated information among different modalities for sentiment analysis.

- By integrating these mathematical formulations into the TETFN architecture, the model effectively combines textual, visual, and acoustic information through attention mechanisms and generates informative labels, enhancing its performance in multimodal sentiment analysis tasks.

  Firgure 3.6 provides an illustration of the TETFN architecture.

3.6.7 LSTM Utterance Level with Multiple Attention Mechanisms

Exploring the Architecture of LSTM with Multiple Levels of Attention Mechanisms. For the LSTM Utterance Level with Multiple Attention Mechanism model, a learning rate of 3×10^{-4}, batch size of 32, and the Adam optimizer were used for training. The model leverages multiple attention mechanisms to enhance the fusion of modality-specific information.

Figure 3.7 offers a depiction of the architecture for LSTM with multiple attention mechanisms.

1. **Input Representation**:

$$x \in \mathbb{R}^{d \times M}, \quad x = [x_1, x_2, ..., x_t, ..., x_M], \quad x_t \in \mathbb{R}^d, \quad t = 0 \text{ to } M$$

2. **LSTM Cell Computation**:

$$X = [h_{t-1}, x_t]$$
$$f_t = \sigma(W_f \cdot X + b_f)$$
$$i_t = \sigma(W_i \cdot X + b_i)$$
$$o_t = \sigma(W_o \cdot X + b_o)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c \cdot X + b_c)$$
$$h_t = o_t \odot \tanh(c_t)$$

3. **Output Representation**:

$$H \in \mathbb{R}^{d \times M}, \quad H = [h_1, h_2, ..., h_t, ..., h_M], \quad h_t \in \mathbb{R}^d$$

4. **Attention Network**:

$$P_t = \tanh(W_h[t] \cdot H)$$
$$\alpha_t = \text{softmax}(w[t]^T \cdot P_t)$$
$$r_t = H \cdot \alpha_t$$

5. **Modified LSTM Representation**:

$$h_t^* = \tanh(W_p[t] \cdot r_t + W_x[t] \cdot h_t), \quad W_p, W_x \in \mathbb{R}^{M \times d \times d}$$

6. **Classification**:

$$Z_t = \text{softmax}((h_t^*)^T \cdot W_{\text{soft}}[t] + b_{\text{soft}}[t]), \quad \hat{y}_t = \arg\max_j (Z_t[j]), \forall j \in \text{class}$$

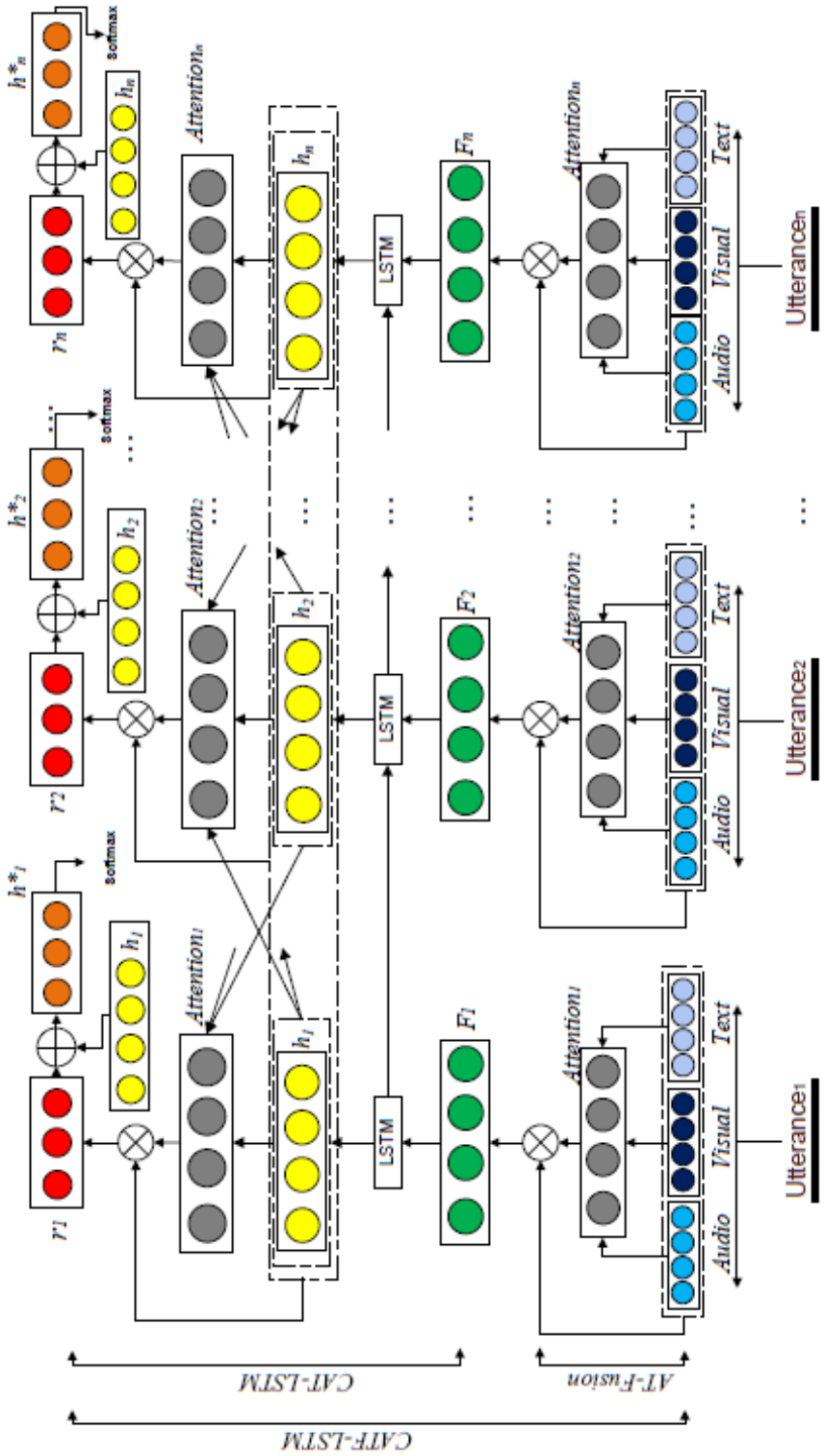

Figure 3.7: CATF-LSTM fuses multimodal inputs via AT-Fusion before feeding them to a CAT-LSTM classifier.

$$W_{\text{soft}} \in \mathbb{R}^{M \times d \times \text{ydim}}, \quad b_{\text{soft}} \in \mathbb{R}^{M \times \text{ydim}}$$

7. **Training**:

$$\text{loss} = -\sum_i \sum_j \log(Z_t[y_j^i]) + \lambda \|\theta\|_2^2$$

Where y is the target class, Z_t is the predicted distribution, λ is the L2 regularization term, and θ is the parameter set.

3.7 Evaluation Metrics

Evaluating classification models is essential for gauging their effectiveness and dependability. A range of metrics is used to measure different facets of a model's performance. This document provides an in-depth explanation of four frequently used metrics: precision, recall, F1 score, accuracy, Loss, MAE, and Corr.

Precision, a metric for measuring the relevance of results, is obtained by dividing the number of true positives by the total number of positive predictions (true positives + false positives):

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

The components involved include:

- **True Positives (TP):** Instances correctly predicted as positive.
- **False Positives (FP):** Instances incorrectly predicted as positive.

High precision indicates that when the model predicts a positive instance, it is likely correct. However, precision does not consider instances the model missed (false negatives).

Precision can be fine-tuned based on the application's requirements. In situations where minimizing false positives is critical, precision becomes a key metric.

Recall (also known as Sensitivity or True Positive Rate) is crucial in scenarios where the impact of false negatives is significant. It is determined by the ratio of true positives to the sum of true positives and false negatives:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The components involved include:

- **True Positives (TP):** Instances correctly predicted as positive.
- **False Negatives (FN):** Instances incorrectly predicted as negative.

High recall means the model effectively captures most positive instances, and it is valuable when identifying all positive instances is critical, even if it results in more false positives.

Recall is particularly important in scenarios where missing positive instances can have significant consequences. For instance, in medical diagnoses, ensuring the identification of all cases of a disease is often more crucial than minimizing false positives.

The **F1 score** represents the harmonic mean of precision and recall, offering a balanced assessment that considers both false positives and false negatives by combining precision and recall into a unified metric.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score falls within the range of 0 to 1, where greater values signify a more optimal balance between precision and recall. This metric proves especially valuable in cases of class imbalances.

The F1 score serves as a holistic measure aiding decision-making when both precision and recall are significant. It excels in situations where maintaining equilibrium between false positives and false negatives holds importance.

Accuracy offers a general gauge of correct predictions but can be deceptive in datasets with imbalances. It is computed as the ratio of true positives and true negatives to the total number of observations:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Observations}}$$

The components involved include:

- **True Positives (TP):** Instances accurately predicted as positive.
- **True Negatives (TN):** Instances correctly identified as negative.
- **False Positives (FP):** Instances mistakenly classified as positive.
- **False Negatives (FN):** Instances mistakenly classified as negative.

Correlation, in sentiment analysis, Corr signifies the correlation coefficient linking the predicted sentiment scores to the actual sentiment labels within your dataset. This metric gauges the strength and orientation of the linear connection between the predicted and actual sentiment values. A high correlation value nearing 1 suggests that the model's predictions strongly correlate with the ground truth labels, indicating proficient performance in capturing the sentiment of the text data.

Loss. Within machine learning models, loss denotes the disparity or deviation between the model's predicted outputs and the factual ground truth labels. This metric measures the efficacy or deficiency of the model's performance throughout the training phase. Generally, the loss function is crafted to diminish this variance by tweaking the model's parameters during training. A decreased loss value denotes superior performance, suggesting that the model's predictions closely align with the actual labels. Loss functions (MSE, cross-entropy, hinge loss) depend on problem type and output format.

MAE, known as Mean Absolute Error (MAE), assesses the model's accuracy in sentiment prediction on a continuous scale. It achieves this by calculating the average of the absolute discrepancies between the predicted sentiment values and the ground truth

While accuracy is straightforward and easily understandable, it may lack depth in scenarios where one class dominates the dataset. In such instances, precision, recall, and the F1 score offer a more nuanced assessment.

In summary, the selection of evaluation metrics depends on the specific objectives and limitations of the application. Precision, recall, and the F1 score provide a more comprehensive insight into a model's performance, especially in cases where certain types of errors carry more significance than others.

3.8 Summary

Chapter 3 elucidates the methodology adopted in the thesis, including an analysis of datasets, architectural details of eight multimodal deep learning approaches, and evaluation metrics.

Initially, the chapter scrutinizes the datasets utilized, namely CMU MOSI and CMU MOSEI, shedding light on their intricacies, challenges, and complexities. This discussion provides a nuanced understanding of the data environment crucial for subsequent analysis.

Subsequently, the chapter delves into the architectural nuances of eight multimodal deep learning approaches. Each approach is meticulously examined, elucidating its design principles, model architectures, and underlying mechanisms, setting the stage for the subsequent evaluation.

Moreover, the chapter expounds on the evaluation metrics employed to assess the performance of multimodal sentiment analysis approaches. It discusses the significance of these metrics in gauging the effectiveness and robustness of the models, ensuring a comprehensive evaluation framework.

Through this multifaceted exploration, Chapter 3 establishes a solid foundation for the subsequent evaluation and analysis of multimodal sentiment analysis, facilitating a thorough understanding of methodologies and evaluation criteria.

CHAPTER 4

Results and Discussion

4.1 Introduction

In this section, we embark on an in-depth examination of the findings resulting from the meticulous evaluation of eight distinct multimodal approaches designed for sentiment analysis using advanced deep learning methodologies. Our evaluation endeavors focus specifically on the MOSI and MOSEI datasets. These datasets have gained prominence within the research community for their rich diversity encompassing textual, audio, and visual modalities, making them ideal platforms for gauging the effectiveness of multimodal sentiment analysis systems under real-life media scenarios.

The multimodal models under scrutiny, namely MISA, MMIM (Multimodal Multi-Instance Model), LSTM W/O Att. (LSTM Without Attention), LMF, SELF MM, LSTM, TETFN, and LSTM, represent a spectrum of innovative approaches aimed at integrating multi-modal information seamlessly. Each model is tailored to capture the intricate sentiment nuances inherent in the complex and diverse content found in the MOSI and MOSEI datasets.

Our thorough assessment involves an in-depth examination of crucial performance metrics such as precision, recall, F1-score, and accuracy. These metrics play a pivotal role as essential yardsticks for evaluating the effectiveness and resilience of multimodal sentiment analysis frameworks across various sentiment categories and diverse datasets. Through this comprehensive analysis, we aim to unveil profound insights into the strengths, limitations, and potential avenues for advancement in multi-modal sentiment analysis using cutting-edge

deep learning methodologies within the challenging and dynamic domains encapsulated by the MOSI and MOSEI datasets.

4.1.1 Performance on MOSI Dataset

As shown in Table 4.1, the detailed results for each approach are presented.

LSTM utterance level multi attention: This approach utilizes LSTM networks with multi-level attention mechanisms for sentiment analysis. The F1 score of 0.73987 illustrates a decent balance between precision and recall in classification, with 73.987% of instances correctly classified. The accuracy of 0.73936 suggests that roughly 73.936% of instances are classified correctly overall. The relatively low loss of 0.01884 indicates efficient model training, signifying successful convergence during the training phase.

LMF on MOSI: LMF is an approach that leverages low-rank weight tensors for the efficient fusion of multimodal features. The relatively high loss of 94.45% and accuracy of 78.28% indicate that this approach performs poorly compared to others. The F1 score of 78.32% suggests that it achieves a moderate balance between precision and recall, However, the overall classification performance falls short of expectations. These findings suggest that LMF may encounter challenges in accurately capturing the underlying patterns in the data, resulting in subpar performance.

MMIM on MOSI: MMIM is a method that maximizes mutual information at both input and fusion levels for multimodal sentiment analysis. The F1 score of 81.72% and accuracy of 81.75% indicate that this approach performs well in classifying sentiment, with approximately 81.72% of instances correctly classified and an overall accuracy of 81.75%. The relatively low loss of 75.11% suggests efficient model training, indicating that the model converges effectively during training.

Self MM on MOSI: Self MM employs self-supervised learning techniques for multimodal sentiment analysis. It attains the top F1 score of 82.72% and an accuracy of 82.8% among all methods, showcasing superior performance in sentiment classification. The low loss of 71.73% suggests efficient model training, with the model converging effectively during training. These results indicate that Self MM effectively leverages self-supervised learning strategies to improve sentiment analysis performance.

TETFN on MOSI: TETFN integrates textual data into audio and visual modalities through

attention mechanisms and transformers. Despite achieving a decent F1 score of 77.7% and an accuracy of 77.75%, the relatively high loss of 94.04% implies less effective model training compared to alternative methods. This suggests potential challenges with model convergence during training, which could impact overall performance.

4.1.2 Evaluation and Discussion

- **Accuracy:** Self MM achieves the highest accuracy, indicating the highest proportion of correctly classified instances. MMIM follows closely with a slightly lower accuracy.
- **F1 Score:** Self MM also secures the top F1 score, representing the harmonic mean of precision and recall and indicating a commendable balance between these two metrics. Following closely, MMIM achieves the second-highest F1 score.
- **Loss:** Self MM has the lowest loss, indicating efficient model training and good convergence. MMIM also has a relatively low loss.
- **MAE:** Self MM has the lowest MAE, indicating the smallest average deviation between predicted and actual sentiment scores.
- **Correlation:** Self MM attains the top correlation coefficient, signifying a robust linear association between predicted and actual sentiment scores.

Overall, Self MM emerges as the top-performing approach on the MOSI dataset, followed closely by MMIM, while LMF exhibits relatively poor performance compared to other methods.

4.1.3 Performance on MOSEI Dataset

As shown in Table 4.2, the detailed results for each approach are presented.

MISA on MOSEI: MISA achieves a well-balanced classification performance, reflected in its F1 score of 0.8426. This is supported by the high precision (0.8451) and recall (0.8415) values, indicating effective instance classification with minimal errors. The 84.15% accuracy suggests a substantial number of correctly classified instances overall. Additionally, the Mean Absolute Error (MAE) of 0.5579 reflects the average absolute difference between predicted and actual sentiment scores

Table 4.1: Table summarizing the outcomes of Multimodal Sentiment Analysis Techniques employed on the MOSI dataset. The findings cover a range of assessment criteria including Accuracy (A), F1 score (F1), Recall (R), Precision (P), Time (T), Loss (L), Mean Absolute Error (MAE), and Correlation (C).

Ref	Model	Method	Fusion Methods Involved	Results (A, F1, R, P, T, L, MAE, C)
[70]	MISA	Involves projecting modalities into modality-invariant and modality-specific subspaces for fusion and analysis.	Combining Text and Image Elements	P(0.82) R(0.81) F1(0.81) A(0.81) MAE(0.87) C(0.71)
[78]	LSTM Utterance Level multi Attn	Contextual attention-based LSTM network for multimodal sentiment classification, prioritizing relevant contextual information for accurate classification	AT-Fusion	L(0.0188) A(0.739) F1(0.739)
[76]	LSTM with Attention	LSTM networks for sequential utterance classification with attention mechanisms to prioritize relevant contextual information	Hierarchical framework integrating context-independent unimodal features with contextual multimodal features	L(0.0264) A(0.7606) F1(0.7611)
[76]	LSTM Without Attention	Involves LSTM networks for sequential utterance classification without attention mechanisms, focusing solely on contextual dependency modeling	Hierarchical framework integrating context-independent unimodal features with contextual multimodal features	L(0.0280) A(0.7527) F1(0.7501)
[96]	SELF MM	Self-supervised multi-task learning strategy utilizing auto-generated unimodal labels, momentum-based update method, and relative distance value for uni-modal supervision in multimodal sentiment analysis	Late fusion approach	L(71.73) F1(82.72) A(82.8) MAE(71.68) C(79.35)
[4]	TETFN	Incorporates textual information into audio and vision modalities using a text-oriented multi-head attention mechanism and cross-modal transformers	Cross-modal attention	L(94.04) F1(77.7) A(77.75) MAE(93.86) C(66.343)
[79]	LMF	Leveraging low-rank weight tensors for efficient multimodal fusion, utilizing modality-specific factors, and computing multimodal representation without explicit tensorization through parallel decomposition	Low-rank Multimodal Fusion	L(94.45) F1(78.32) A(78.28) MAE(94.28) C(65.97)
[3]	MMIM	Hierarchical mutual information maximization for multimodal sentiment analysis, integrating input and fusion-level MI maximization, parametric learning, non-parametric Gaussian Mixture Models, and comprehensive experiments to achieve competitive results	MI maximization	L(75.11) F1(81.72) A(81.75) MAE(75.21) C(77.01)

Table 4.2: Table summarizing the outcomes of Multimodal Sentiment Analysis Techniques employed on the MOSEI dataset. The findings cover a range of assessment criteria including Accuracy (A), F1 score (F1), Recall (R), Precision (P), Time (T), Loss (L), Mean Absolute Error (MAE), and Correlation (C).

Ref	Model	Method	Fusion Methods Involved	Results (A, F1, R, P, T, L, MAE, C)
[70]	MISA	involves projecting modalities for fusion and analysis into subspaces that are both modality-specific and modality-invariant.	Combining Text and Image Elements	P(0.8451) R(0.8415) F1(0.8426) A(0.8415) MAE(0.5579) Corr(0.7433)
[78]	LSTM Utterance Level multi Attn	Contextual attention-based LSTM network for multimodal sentiment classification, prioritizing relevant contextual information for accurate classification	AT-Fusion	L(0.00137) A(0.59428) F1(0.579263)
[76]	LSTM with Attention	LSTM networks for sequential utterance classification with attention mechanisms to prioritize relevant contextual information	Hierarchical framework integrating context-independent unimodal features with contextual multimodal features	L(0.00138) A(0.59341) F1(0.580733)
[76]	LSTM Without Attention	Involves LSTM networks for sequential utterance classification without attention mechanisms, focusing solely on contextual dependency modeling	Hierarchical framework integrating context-independent unimodal features with contextual multimodal features	L(0.001414) A(0.58691) F1(0.578075)
[96]	SELF MM	Multimodal sentiment analysis approach using a self-supervised learning framework with automatically generated labels for each modality. This framework incorporates a momentum-based update strategy and leverages relative distance information to enhance uni-modal supervision	Late fusion approach	F1(85.3) A(85.17) MAE(0.53) C(0.765)
[4]	TETFN	Employs a text-guided multi-head attention mechanism within a cross-domain transformer framework to facilitate the fusion of textual information with audio and visual data	Cross-modal attention	F1(85.27) A(85.18) MAE(0.551) C(0.748)
[79]	LMF	Leveraging low-rank weight tensors for efficient multimodal fusion, utilizing modality-specific factors, and computing multimodal representation without explicit tensorization through parallel decomposition	Low-rank Multimodal Fusion	C(0.677) F1(82.1) A(82) MAE(0.623)
[3]	MMIM	Optimizes sentiment information flow in multimodal analysis through hierarchical mutual dependence maximization., parametric learning, non-parametric Gaussian Mixture Models, and comprehensive experiments to achieve competitive results	MI maximization	F1(82) A(82.24) MAE(0.526) C(0.772)

LSTM without ATT, LSTM with ATT, and LSTM utterance level multi ATT on MOSEI: These LSTM-based models perform similarly in terms of loss, accuracy, and F1 score. They achieve relatively low losses, indicating efficient model training. The F1 scores suggest moderate performance in sentiment classification. The accuracies are above 58

LMF on MOSEI: LMF performs relatively poorly compared to other methods. The high loss and relatively low accuracy and F1 score indicate suboptimal performance. The correlation coefficient indicates a moderate linear connection between predicted and actual sentiment scores. However, the relatively high MAE implies larger deviations between predicted and actual scores.

MMIM on MOSEI: MMIM achieves a high F1 score and accuracy, indicating effective sentiment classification. The relatively low loss and MAE suggest efficient model training and small deviations between predicted and actual scores.The strong correlation coefficient signifies a robust linear connection between predicted and actual sentiment scores.

Self MM on MOSEI: Self MM emerges as the top-performing approach on the MOSEI dataset. It achieves the highest F1 score and accuracy among all methods, indicating superior performance in sentiment classification. The low loss and MAE suggest efficient model training and small deviations between predicted and actual scores. The elevated correlation coefficient suggests a robust linear association between predicted and actual sentiment scores.

TETFN on MOSEI: TETFN performs comparably to Self MM, achieving a high F1 score and accuracy. However, it exhibits a slightly higher MAE and lower correlation coefficient compared to Self MM. Overall, TETFN demonstrates effective sentiment classification performance on the MOSEI dataset.

Evaluation and Discussion

- **Accuracy:** Self MM achieves the highest accuracy, indicating the highest proportion of correctly classified instances, followed closely by TETFN.
- **F1 Score:** The Self MM model also achieves the highest F1 score, indicating a commendable balance between precision and recall, closely followed by TETFN.
- **MAE:** MMIM has the lowest MAE, indicating smaller average deviations between predicted and actual sentiment scores.

- **Correlation:** The Self MM model achieves the highest correlation coefficient, indicating a strong linear correlation between predicted and actual sentiment scores.

In summary, Self MM stands out as the leading approach on the MOSEI dataset, with TETFN and MMIM following closely behind, whereas LMF shows comparatively weaker performance among the methods evaluated. These results highlight the effectiveness of self-supervised learning methods in improving sentiment analysis performance on multimodal datasets like MOSEI.

4.2 Discussion

The assessment of different models aimed to gauge their performance on the specified task. This summary discusses the key observations regarding the results, **noting that certain metrics were unavailable due to their absence in the code of the respective multimodal models**:

Table 4.3: Results on MOSI Dataset

Model	**Acc**	**F1**	**Loss**	**MAE**	**Corr**
MISA	0.81	0.81	-	0.87	0.71
LSTM multi level attention	0.74	0.74	0.0188	-	-
LSTM with attention	0.76	0.76	0.0264	-	-
LSTM without attention	0.75	0.75	0.0280	-	-
SELF MM	0.83	0.83	-	0.71	0.79
TETFN	0.78	0.78	0.94	0.93	0.66
LMF	0.78	0.78	0.94	0.94	0.65
MMIM	0.82	0.82	0.75	0.75	0.77

Table 4.4: Results on MOSEI Dataset

Model	**Acc**	**F1**	**Loss**	**MAE**	**Corr**
MISA	0.84	0.84	-	0.55	0.74
LSTM multi level attention	0.59	0.57	0.0013	-	-
LSTM with attention	0.59	0.58	0.0013	-	-
LSTM without attention	0.58	0.57	0.00141	-	-
SELF MM	0.85	0.85	-	0.53	0.76
TETFN	0.85	0.85	-	0.551	0.74
LMF	0.82	0.82	0.677	0.62	-
MMIM	0.82	0.82	-	0.52	0.77

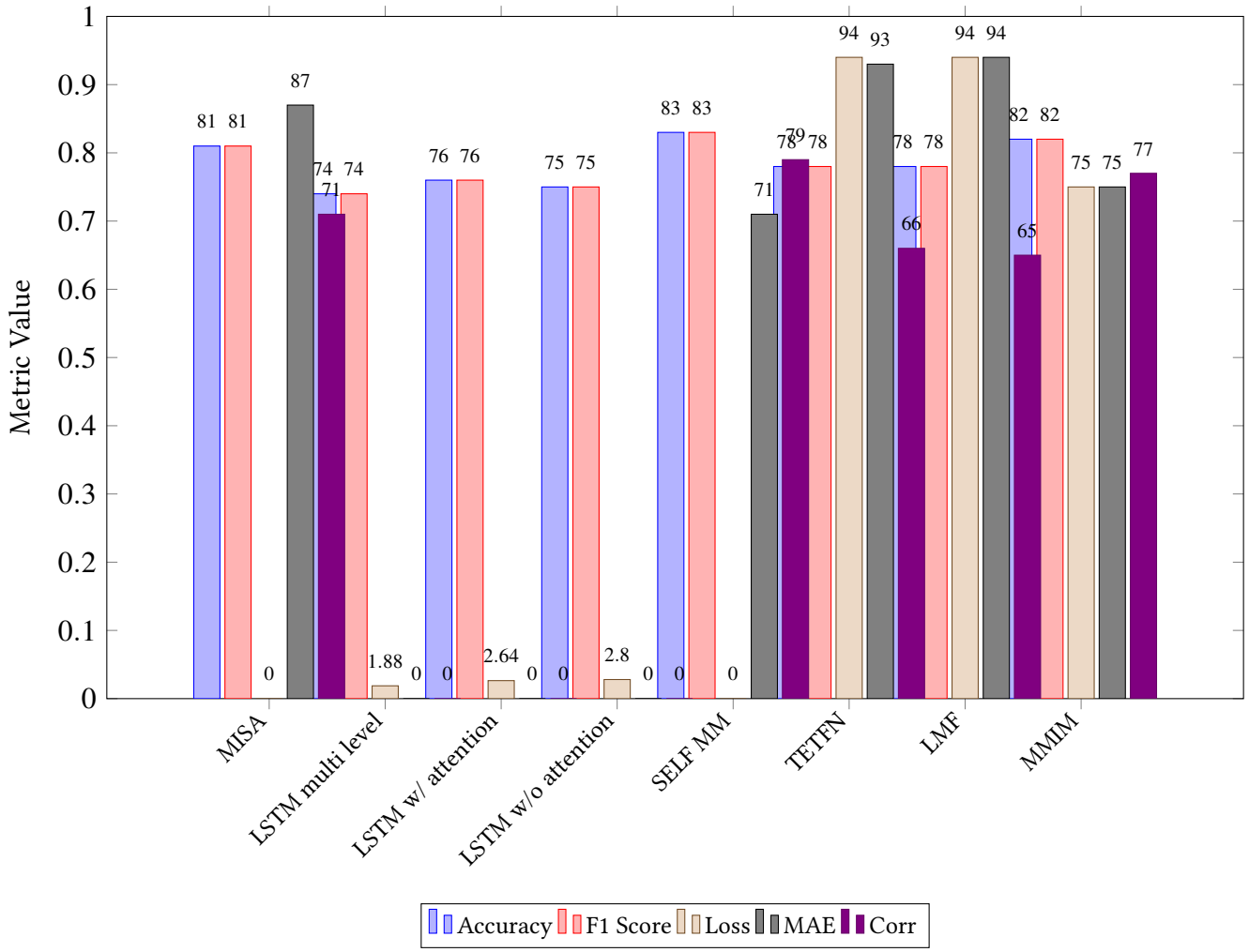

Figure 4.1: Model Comparison on MOSI Dataset Across Metrics

4.2.1 Bar Graph Discussion for MOSI

The grouped bar chart 4.1 provides a clear visual comparison of the performance of various models on the MOSI dataset across multiple metrics, including Accuracy, F1 Score, Loss, Mean Absolute Error (MAE), and Correlation (Corr).

Accuracy and F1 Score: Most models exhibit similar performance in terms of accuracy and F1 score, with values ranging from 0.74 to 0.83. Notably, the SELF MM model achieves the highest performance in both metrics, closely followed by MMIM and MISA, indicating their robustness in classification tasks on the MOSI dataset.

Loss: The Loss metric varies significantly among the models, with the TETFN and LMF models showing higher loss values, suggesting these models may struggle with overfitting or optimization issues. In contrast, the MMIM model demonstrates relatively low loss, indicating better generalization.

MAE: The Mean Absolute Error metric highlights that MISA and SELF MM have lower error rates, suggesting better accuracy in continuous prediction tasks. TETFN and LMF models, however, have higher MAE values, which could reflect challenges in their ability to predict finer-grained sentiment values accurately.

Correlation: The correlation metric shows how well the models' predictions correlate with the ground truth. SELF MM and MMIM lead in this metric, indicating that these models not only achieve high classification accuracy but also maintain strong alignment with the true sentiment scores.

Overall, SELF MM and MMIM emerge as the most consistent performers across the various metrics, making them strong candidates for applications requiring both classification and regression tasks in sentiment analysis. The visualization effectively highlights the strengths and weaknesses of each model, facilitating a more informed selection based on specific performance criteria.

4.2.2 Bar Graph discussion for Mosei

The grouped bar chart 4.2 illustrates the performance of various models on the MOSEI dataset across multiple metrics: Accuracy, F1 Score, Loss, Mean Absolute Error (MAE), and Correlation (Corr).

Accuracy and F1 Score: The SELF MM and TETFN models exhibit the highest accuracy

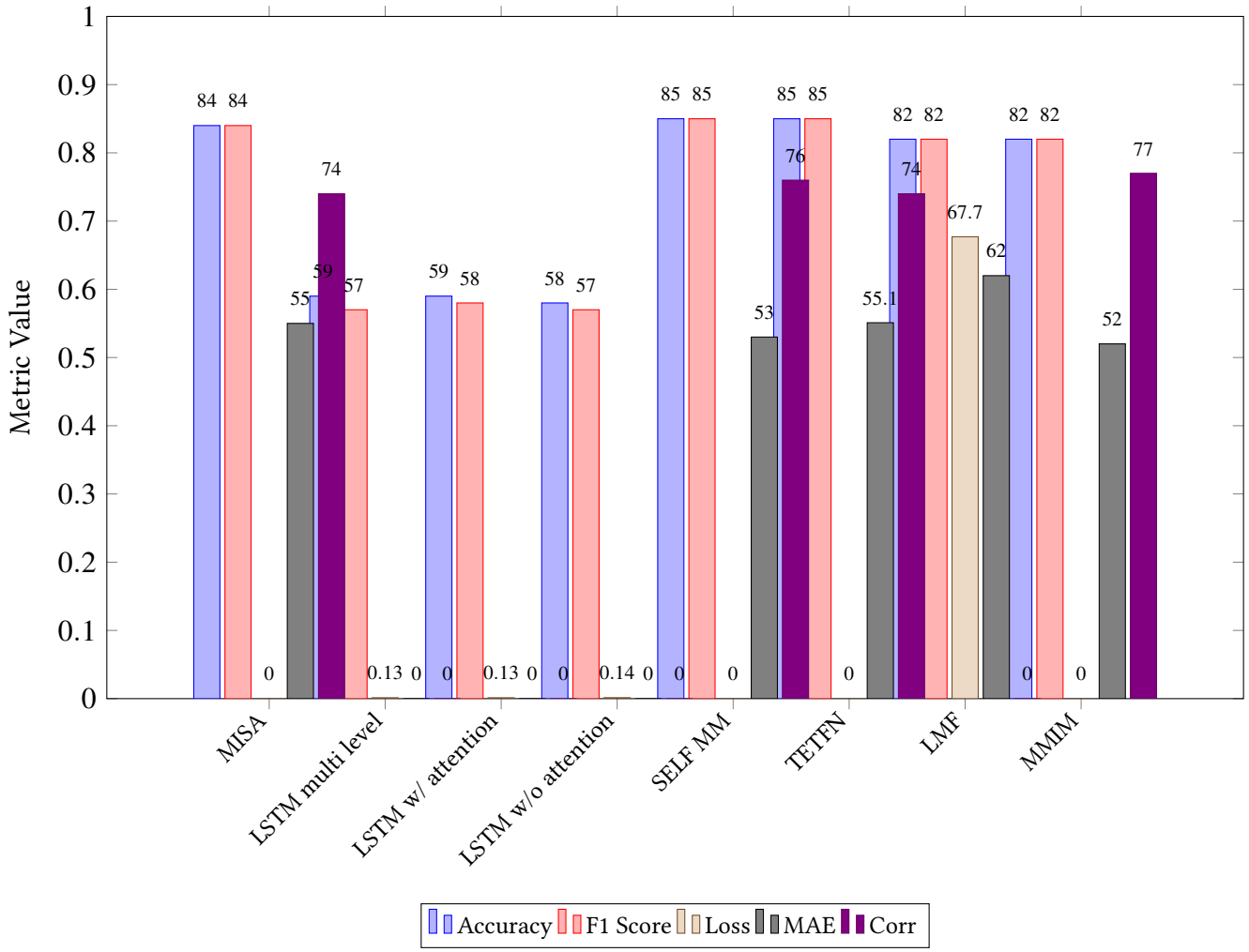

Figure 4.2: Model Comparison on MOSEI Dataset Across Metrics

and F1 scores, both reaching 0.85, indicating their strong performance in classification tasks on the MOSEI dataset. MISA follows closely with an accuracy and F1 score of 0.84, showing its robustness. The LSTM-based models demonstrate lower accuracy and F1 scores, around 0.58 to 0.59, suggesting these models may struggle with the complexity of the MOSEI dataset.

Loss: Loss values vary significantly, with the LMF model showing the highest loss at 0.677, indicating potential optimization issues or overfitting. The other models either have no loss value reported or display minimal loss, which suggests better training stability for those models.

MAE: In terms of Mean Absolute Error, the MMIM model achieves the lowest MAE at 0.52, reflecting its strong prediction accuracy in continuous sentiment analysis tasks. The LMF model shows a relatively high MAE of 0.62, indicating a less accurate prediction of sentiment scores.

Correlation: The correlation metric is crucial for understanding how well the model's predictions align with the ground truth. MMIM leads with a correlation of 0.77, followed by SELF MM at 0.76, highlighting their effectiveness in capturing the nuances of the MOSEI dataset.

Overall, the SELF MM, TETFN, and MMIM models stand out as the top performers across most metrics, making them well-suited for sentiment analysis tasks on the MOSEI dataset. This visualization helps identify the strengths and weaknesses of each model, providing valuable insights for selecting the appropriate model based on specific performance criteria.

4.2.3 Key Observations

After analyzing the outcomes from various models on both the MOSI and MOSEI datasets (refer to Tables 4.3 and 4.4), several crucial observations surface, shedding light on the efficacy and subtleties of each method. **It's noteworthy that certain metrics were not attainable due to their absence in the code of the respective multimodal models.**

- **MISA Model:** The Multimodal Integrated Sentiment Analysis (MISA) model showcases outstanding performance across both datasets. With high accuracy and F1 scores, it demonstrates robust sentiment classification capabilities. Additionally, MISA exhibits minimal Mean Absolute Error (MAE), indicating precise sentiment intensity prediction.

Its moderate correlation with actual values suggests reliable sentiment analysis across modalities, leveraging integrated features effectively.

- **LSTM Models:** The LSTM-based models, including multi-level attention variants, demonstrate moderate performance on both datasets. While they exhibit reasonable accuracy and F1 scores, there is room for improvement, particularly in optimizing convergence during training. Despite slightly higher loss values, these models present competent sentiment classification capabilities.

- **SELF MM Model:** The Self-Modulating Multimodal (SELF MM) model emerges as a top performer, surpassing others in accuracy and F1 scores. Notably, it excels in predicting sentiment intensity, with low MAE and high correlation values. This underscores its efficacy in capturing subtle sentiment nuances across modalities.

- **Transformer-based Models:** Transformer-based architectures, such as TETFN and LMF, exhibit comparable performance across datasets. While they demonstrate moderate accuracy and F1 scores, exploring alternative fusion strategies may further enhance their effectiveness in sentiment analysis tasks.

- **MMIM Model:** The Multimodal Interaction Model (MMIM) shows promising outcomes, especially on the MOSI dataset. Boasting high accuracy and F1 scores, it comes in second only to SELF MM. Its inclusion of interaction-aware features enables sophisticated sentiment analysis, demonstrating its potential for multimodal tasks.

These comprehensive observations provide valuable insights into each model's performance and characteristics, paving the way for future improvements in multimodal sentiment analysis methodologies.

4.3 Summary

The thesis embarks on an extensive comparative inquiry into a spectrum of multimodal sentiment analysis (MSA) frameworks, aiming to elucidate their effectiveness, contributions, and potential implications in sentiment analysis tasks. Commencing with MISA, the focus is on

the pivotal role of multimodal representation learning, particularly in amalgamating disparate information streams encompassing textual, speech, and visual modalities. MISA's intricate fusion mechanisms, complemented by mathematical methodologies such as Central Moment Discrepancy (CMD) and soft orthogonality constraints, underscore its superior efficacy in discerning affective states within diverse multimedia landscapes.

Next, the analysis delves into the MultiModal InfoMax (MMIM) framework, praised for its ability to discern redundant information and capture invariant trends across modalities by maximizing mutual information (MI). This blended approach, combining parametric and non-parametric methods for MI estimation, enhances MMIM's relevance in real-world scenarios. Similarly, the LSTM W/O Attn framework introduces a new direction in MSA by utilizing Long Short-Term Memory networks to grasp contextual dependencies among sequential utterances within videos, eliminating the need for attention mechanisms. The integration of multimodal features, facilitated through both non-hierarchical and hierarchical frameworks, results in subtle improvements in sentiment analysis performance.

Furthermore, the LMF method introduces a scalable solution for integrating multimodal information effectively, showcasing robust performance across various rank settings. SELF MM presents an innovative self-supervised multi-task learning strategy, harnessing auto-generated unimodal labels to bolster stability and reliability in sentiment analysis endeavors. The TETFN tackles MSA challenges by incorporating textual information to enrich fusion representations, attaining superior performance through its text-enhanced transformer module.

Lastly, the emergence of the Contextual Attention-based LSTM (CAT-LSTM) network with utterance-level multiple attention heralds a promising paradigm for MSA. Its adeptness in capturing contextual cues and prioritizing pertinent information for sentiment analysis signifies a notable advancement. Through rigorous experimental validation and comparative scrutiny, the thesis furnishes profound insights into the merits and demerits of each method, thereby laying a robust groundwork for the evolution of multimodal sentiment analysis methodologies.

Further exploration could delve into the optimization of fusion mechanisms, exploration of novel attention mechanisms, investigation of domain adaptation techniques, and application of multimodal sentiment analysis in real-world scenarios, thereby enriching the understanding

and practical utility of MSA frameworks across diverse domains and contexts.

CHAPTER 5

Conclusion

5.1 Summary of Contributions

In this segment, we provide an outline the unique aspects of our performance evaluation research in Multimodal Sentiment Analysis, setting it apart from existing academic literature. Past surveys on multimodal analysis in text, audio, and videos [97, 98, 99, 100, 101] predominant focus has been on particular domains such as document-based, feature-based, and visual sentiment analysis. However, recent academic inquiries have revealed limited exploration into the thorough comparison of these domains within a single study. Furthermore, there exists a dearth of exhaustive performance evaluations specifically devoted to Multimodal Sentiment Analysis (MSA), inclusive of detailed deliberations on deep learning models. Hence, our research endeavors to address this gap by conducting a meticulous performance evaluation and drawing upon pertinent academic literature. The noteworthy contributions of this performance evaluation study are outlined below.

Comprehensive Modalities Coverage: Our study comprehensively evaluates the efficacy of eight multimodal approaches across diverse modalities, furnishing a holistic appraisal of their effectiveness.

Examination of Application Areas: In contrast to prior research, which has been compartmentalized into specific domains, our study probes the performance of multimodal approaches across a spectrum of application areas within MSA.

Detailed Evaluation Metrics: We employ a diverse range of evaluation metrics to metic-

ulously assess the performance of each multimodal approach, affording insights into their respective strengths and weaknesses.

Analysis of Deep Learning Models: The study conducts an in-depth analysis of the performance of deep learning models integrated into the evaluated multimodal approaches, offering valuable insights into their efficacy in MSA tasks.

Comparison of Fusion Methods: Various fusion techniques utilized to integrate multiple modalities in MSA are scrutinized, elucidating their relative advantages and drawbacks in terms of performance.

Dataset Details and Architectures: Section 3 provides intricate descriptions of the datasets utilized and delineates the architectures of the eight multimodal approaches assessed in our study.

Results and Discussion: Chapter 4 delves into the key observations derived from the evaluation of the eight multimodal approaches, elucidating their performance and effectiveness, shedding light on their respective strengths and weaknesses. Subsequently, Chapter 5 encapsulates the culmination of our research endeavor. It presents the conclusion drawn from the findings, acknowledging the contributions made and discussing the implications for future research in MSA. Additionally, Chapter 5 addresses the limitations encountered during the study, providing insights into areas for improvement and potential avenues for future research exploration.

Through this comprehensive performance evaluation, our study aims to contribute to scholarly understanding of effective multimodal approaches in MSA, offering valuable insights for future research endeavors in this domain.

5.1.1 Answers to Research Questions

RQ1: Which multimodal sentiment analysis models achieve the highest performance across critical evaluation metrics such as accuracy, precision, recall, and F1-score when applied to the challenging sentiment analysis tasks posed by the MOSI and MOSEI datasets? What insights can be derived from their comparative analysis based on these metrics?

Upon analyzing the MOSI dataset, it was found that the "SELF MM" model demonstrated the greatest accuracy, achieving a score of 0.83, along with an F1-score of 0.83. Meanwhile, on the MOSEI dataset, both the "SELF MM" and "TETFN" models exhibited the highest accuracy

and F1-score, each reaching 0.85.

The "SELF MM" model consistently performs well on both MOSI and MOSEI datasets, achieving high accuracy and F1-scores. However, other models such as "TETFN" also perform competitively. Insights from the comparative analysis suggest that certain models may excel on specific datasets, highlighting the importance of dataset-specific model evaluation.

RQ2: What are the respective strengths and weaknesses of each multimodal sentiment analysis model in effectively capturing and interpreting nuanced sentiments across diverse modalities, including text, audio, and visual cues, measured in terms of precision, recall, F1-score, and accuracy?

The "SELF MM" model demonstrates strength in capturing nuanced sentiments across diverse modalities, achieving high accuracy and F1-scores on both MOSI and MOSEI datasets. However, other models like "TETFN" also exhibit competitive performance. Weaknesses may vary depending on the specific model and dataset characteristics, requiring further analysis.

RQ3: How can the performance of the multimodal sentiment analysis models be further optimized or enhanced, and what novel strategies or modifications can be proposed based on the evaluation results, particularly in the context of precision, recall, F1-score, and accuracy?.

Based on the evaluation results, further optimization of the models could involve fine-tuning model architectures, incorporating additional features, or exploring ensemble methods to improve performance metrics such as precision, recall, F1-score, and accuracy. Novel strategies may include leveraging transfer learning techniques or exploring advanced multimodal fusion approaches to enhance model robustness and generalization capabilities.

These responses offer valuable insights into the effectiveness and potential areas for enhancement of the applied multimodal sentiment analysis models. Additional examination and experimentation might be required to fine-tune and optimize the models to suit particular applications and datasets.

5.2 Future Work

In this study, which compares the results of eight multimodal approaches integrating audio, text, and visual data on the MOSEI and MOSI datasets, several avenues for future research emerge.

Firstly, there is a need for optimizing multimodal architectures, including fine-tuning existing models with attention mechanisms or transformer-based architectures, and leveraging complementary information from each modality more effectively.

Furthermore, exploring advanced feature engineering methods and representation learning techniques is essential for capturing nuanced patterns and temporal dependencies across modalities. Investigating domain adaptation and transfer learning strategies is also necessary to enhance model generalization, particularly in new domains or tasks with limited labeled data.

Incorporating background knowledge and common-sense reasoning into models can enhance semantic understanding and contextual modeling. Moreover, prioritizing interpretability and explainability by integrating human-readable explanations and visualizations can aid comprehension and trust in model decisions.

Conducting extensive cross-modal fusion and interaction modeling experiments, along with real-world applications and user studies, will provide insights into practical utility and effectiveness. Lastly, emphasis on benchmarking and comparative analysis remains essential for evaluating model performance rigorously and contributing to the development of standardized evaluation metrics.

In conclusion, the future of multimodal sentiment analysis holds promising avenues for research and development. By focusing on optimizing multimodal architectures, exploring advanced feature engineering methods, and investigating domain adaptation strategies, we can enhance the effectiveness and generalization capabilities of multimodal models. Moreover, emphasizing interpretability, conducting extensive experiments, and performing rigorous benchmarking will drive the field forward and aid in the development of standardized evaluation metrics. As researchers continue to push the boundaries of multimodal sentiment analysis, We can anticipate significant advancements in understanding and utilizing the complex interactions between audio, text, and visual modalities for sentiment analysis across various domains and applications.

Statement of Conflict of Interest

The authors declare that they have no conflicts of interest. This study was conducted impartially, and the findings presented in this paper are solely based on the data and analysis performed during the

research phase. There are no financial or other relationships that could be perceived as conflicting interests.

Bibliography

[1] K. Yang, H. Xu, and K. Gao, "Cm-bert: Cross-modal bert for text-audio sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 521–528, 2020.

→ [pxv], [p18]

[2] A. Pandey and D. K. Vishwakarma, "Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey," *Applied Soft Computing*, p. 111206, 2023.

→ [pxv], [p22], [p24]

[3] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 9180–9192, Association for Computational Linguistics, Nov. 2021. → [pxv], [p30], [p31], [p51], [p74], [p75]

[4] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "Tetfn: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

→ [pxv], [p32], [p62], [p74], [p75]

[5] N. Mittal, D. Sharma, and M. L. Joshi, "Image sentiment analysis using deep learning," in *2018 IEEE/WIC/ACM international conference on web intelligence (WI)*, pp. 684–687, IEEE, 2018.

→ [p1]

[6] B. Seetharamulu, B. N. K. Reddy, and K. B. Naidu, "Deep learning for sentiment analysis based on customer reviews," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5, IEEE, 2020. → [p1]

[7] S. Wen, H. Wei, Y. Yang, Z. Guo, Z. Zeng, T. Huang, and Y. Chen, "Memristive lstm network for sentiment analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 3, pp. 1794–1804, 2019. → [p1]

[8] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American society for information science and technology*, vol. 60, no. 11, pp. 2169–2188, 2009. → [p1]

[9] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1275–1284, 2009. → [p1]

[10] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data*, pp. 415–463, Springer, 2012. → [p1]

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," *arXiv preprint cs/0205070*, 2002. → [p12]

[12] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*, pp. 231–240, 2008. → [p12]

[13] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011. → [p12]

[14] F. Sebastiani and A. Esuli, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proceedings of the 5th international conference on language resources and evaluation*, pp. 417–422, European Language Resources Association (ELRA) Genoa, Italy, 2006. → [p12]

[15] M. Hearst, "Direction-based text interpretation as an information access refinement text-based intelligent systems, l," *Text-based intelligent systems: Current research and practice in information extraction and retrieval*, pp. 257–274, 1992. → [p13]

[16] W. Sack, "On the computation of point of view," in *AAAI*, vol. 1488, 1994. → [p13]

[17] B. Pang, L. Lee, *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008. → [p13]

[18] P. Nakov, T. Zesch, D. Cer, and D. Jurgens, "Proceedings of the 9th international workshop on semantic evaluation (semeval 2015)," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015. → [p13]

[19] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," *arXiv preprint arXiv:1308.6242*, 2013. → [p13], [p15]

[20] Y. Miura, S. Sakaki, K. Hattori, and T. Ohkuma, "Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 628–632, 2014. → [p13]

[21] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Webis: An ensemble for twitter sentiment detection," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 582–589, 2015. → [p13], [p15]

[22] J. M. Deriu, M. Gonzenbach, F. Uzdilli, A. Lucchi, V. De Luca, and M. Jaggi, "Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 1124–1128, 2016. → [p13], [p15]

[23] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Computational linguistics*, vol. 16, no. 2, pp. 79–85, 1990. → [p13]

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. → [p13], [p27]

[25] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, p. 2, 2019. → [p14]

[26] Y. Zhao, E. Soerjodjojo, and H. Che, "Methods to enhance bert in aspect-based sentiment classification," in *2022 Euro-Asia Conference on Frontiers of Computer Science and Information Technology (FCSIT)*, pp. 21–27, 2022. → [p15]

[27] A. H. Ahmed Abbasi and M. Dhar, "Benchmarking twitter sentiment analysis tools," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, may. European Language Resources Association (ELRA)*, 2014. → [p15]

[28] A. Hassan, A. Abbasi, and D. Zeng, "Twitter sentiment analysis: A bootstrap ensemble framework," in *2013 international conference on social computing*, pp. 357–364, IEEE, 2013. → [p15]

[29] Q.-T. Truong and H. W. Lauw, "Visual sentiment analysis for review images with item-oriented and user-oriented cnn," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1274–1282, 2017. → [p16]

[30] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009. → [p16]

[31] W. Di, N. Sundaresan, R. Piramuthu, and A. Bhardwaj, "Is a picture really worth a thousand words? -on the role of images in e-commerce," in *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 633–642, 2014. → [p16]

[32] A. Goswami, N. Chittar, and C. H. Sung, "A study on the impact of product images on user clicks for online shopping," in *Proceedings of the 20th international conference companion on World wide web*, pp. 45–46, 2011. → [p16]

[33] R. He, C. Lin, J. Wang, and J. McAuley, "Sherlock: sparse hierarchical embeddings for visually-aware one-class collaborative filtering," *arXiv preprint arXiv:1604.05813*, 2016. → [p16]

[34] R. He and J. McAuley, "Vbpr: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016. → [p16]

[35] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pp. 105–112, 2013. → [p16]

[36] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 944–958, 2014. → [p16]

[37] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Copula ordinal regression for joint estimation of facial action unit intensity," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 4902–4910, 2016. → [p16]

[38] S. Kaltwang, S. Todorovic, and M. Pantic, "Doubly sparse relevance vector machine for continuous facial behavior estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 9, pp. 1748–1761, 2015. → [p16]

[39] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pp. 468–475, IEEE, 2017. → [p16]

[40] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, pp. 169–176, 2011. → [p17], [p25], [p37]

[41] M. A. Nicolaou, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1522–1526, IEEE, 2014. → [p17]

[42] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust correlated and individual component analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1665–1678, 2015. → [p17]

[43] W. Wei-Ning, Y. Ying-Lin, and J. Sheng-Ming, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3534–3539, IEEE, 2006. → [p17]

[44] V. Yanulevskaya, J. C. van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *2008 15th IEEE international conference on Image Processing*, pp. 101–104, IEEE, 2008. → [p17]

[45] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 715–718, 2010. → [p17]

[46] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *Proceedings of the second international workshop on issues of sentiment discovery and opinion mining*, pp. 1–8, 2013. → [p17]

[47] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 223–232, 2013. → [p17]

[48] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014. → [p17]

[49] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 29, 2015. → [p17]

[50] V. Campos, B. Jou, and X. Giro-i Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017. → [p17]

[51] V. Campos, A. Salvador, X. Giró-i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction," in *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pp. 57–62, 2015. → [p17]

[52] T. Narihira, D. Borth, S. X. Yu, K. Ni, and T. Darrell, "Mapping images to sentiment adjective noun pairs with factorized neural nets," *arXiv preprint arXiv:1511.06838*, 2015. → [p17]

[53] B. Jou and S.-F. Chang, "Deep cross residual learning for multitask visual recognition," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 998–1007, 2016. → [p17]

[54] A. Mathews, L. Xie, and X. He, "Senticap: Generating image descriptions with sentiments," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016. → [p17]

[55] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3, pp. 1970–1973, IEEE, 1996. → [p17]

[56] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011. → [p17]

[57] S. Crouch and R. Khosla, "Sentiment analysis of speech prosody for dialogue adaptation in a diet suggestion program," *ACM SIGHIT Record*, vol. 2, no. 1, pp. 8–8, 2012. → [p17]

[58] F. Mairesse, J. Polifroni, and G. Di Fabbrizio, "Can prosody inform sentiment analysis? experiments on short spoken reviews," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5093–5096, IEEE, 2012. → [p18], [p25]

[59] J. Pereira, J. Luque, and X. Anguera, "Sentiment retrieval on web reviews using spontaneous natural speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4583–4587, IEEE, 2014. → [p18]

[60] V. Pérez-Rosas and R. Mihalcea, "Sentiment analysis of online spoken reviews.," in *INTERSPEECH*, pp. 862–866, 2013. → [p18]

[61] L. Kaushik, A. Sangwan, and J. H. Hansen, "Sentiment extraction from natural audio streams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8485–8489, IEEE, 2013. → [p18]

[62] L. Kaushik, A. Sangwan, and J. H. Hansen, "Automatic sentiment extraction from youtube videos," in *2013 IEEE Workshop on automatic speech recognition and understanding*, pp. 239–244, IEEE, 2013. → [p18]

[63] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion recognition using imperfect speech recognition," 2010. → [p18]

[64] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018. → [p21]

[65] W. De Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language*, vol. 30, no. 1, pp. 61–98, 2015. → [p21]

[66] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019. → [p21]

[67] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990. → [p21]

[68] P. Netrapalli, "Stochastic gradient descent and its variants in machine learning," *Journal of the Indian Institute of Science*, vol. 99, no. 2, pp. 201–213, 2019. → [p21]

[69] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014. → [p21]

[70] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020. → [p24], [p74], [p75]

[71] S. Somasundaran, J. Wiebe, P. Hoffmann, and D. Litman, "Manual annotation of opinion categories in meetings," in *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, pp. 54–61, 2006. → [p25]

[72] S. Raaijmakers, K. P. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 466–474, 2008. → [p25]

[73] F. Metze, T. Polzehl, and M. Wagner, "Fusion of acoustic and linguistic features for emotion detection," in *2009 IEEE International Conference on Semantic Computing*, pp. 153–160, IEEE, 2009. → [p25]

[74] V. P. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013. → [p25], [p37]

[75] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2539–2544, 2015. → [p26]

[76] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883, 2017. → [p26], [p35], [p74], [p75]

[77] D. Olson, "From utterance to text: The bias of language in speech and writing," *Harvard educational review*, vol. 47, no. 3, pp. 257–281, 1977. → [p27]

[78] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 1033–1038, IEEE, 2017. → [p27], [p33], [p74], [p75]

[79] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018. → [p28], [p56], [p74], [p75]

[80] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat.*, vol. 1, pp. 397–401, IEEE, 1997. → [p29]

[81] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 366–371, IEEE, 1998. → [p29]

[82] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013. → [p29], [p37]

[83] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," *arXiv preprint arXiv:1905.05812*, 2019. → [p31]

[84] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016. → [p31]

[85] J. Yuan, M. Liberman, *et al.*, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008. → [p31]

[86] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018. → [p32]

[87] Q. Shi, J. Fan, Z. Wang, and Z. Zhang, "Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain," *Pattern Recognition*, vol. 130, p. 108837, 2022. → [p33]

[88] M. Angelou, V. Solachidis, N. Vretos, and P. Daras, "Graph-based multimodal fusion with metric learning for multimodal classification," *Pattern Recognition*, vol. 95, pp. 296–307, 2019. → [p33]

[89] Y. Liu, L. Liu, Y. Guo, and M. S. Lew, "Learning visual and textual representations for multimodal matching and classification," *Pattern Recognition*, vol. 84, pp. 51–67, 2018. → [p33]

[90] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5301–5311, 2021. → [p34]

[91] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. → [p35]

[92] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008. → [p36]

[93] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016. → [p38]

[94] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. → [p38]

[95] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *Ieee Access*, vol. 7, pp. 63373–63394, 2019. → [p48]

[96] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 10790–10797, May 2021. → [p74], [p75]

[97] R. Kaur and S. Kautish, "Multimodal sentiment analysis: A survey and comparison," *Research anthology on implementing sentiment analysis across multiple disciplines*, pp. 1846–1870, 2022. → [p85]

[98] G. Chandrasekaran, T. N. Nguyen, and J. Hemanth D, "Multimodal sentimental analysis for social media applications: A comprehensive review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, p. e1415, 2021. → [p85]

[99] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, pp. 617–663, 2019. → [p85]

[100] R. Ji, D. Cao, Y. Zhou, and F. Chen, "Survey of visual sentiment prediction for social media analysis," *Frontiers of Computer Science*, vol. 10, pp. 602–611, 2016. → [p85]

[101] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, 2022. → [p85]