**BUILDING A MULTIMODAL SENTIMENT ANALYSIS SYSTEM THAT COMBINES TEXTUAL, AUDIO AND VISUAL CUES.**

**BY**

**DINA ANJOLAOLUWA AFOLASHADE.**

**MATRIC NO: 21/8828.**

**A PROJECT WRITTEN AND SUBMITTED TO THE DEPARTMENT OF**

**COMPUTER SCIENCE, COLLEGE OF PURE AND APPLIED SCIENCES (COPAS),**

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF**

**BACHELOR OF SCIENCE (B.Sc.) DEGREE IN COMPUTER SCIENCE OF CALEB**

**UNIVERSITY, LAGOS.**

**MAY, 2025.**

**DECLARATION**

I, **DINA ANJOLAOLUWA AFOLASHADE**, with matriculation number **21/8828** of the

Department of Computer Science, do hereby declare that this project is entirely my work and

composition. The work embodied in this project has not been submitted in candidature for

any degree and is not concurrently being submitted for any other degree. All references made

to works of other persons have been duly acknowledged.

**Signature**………………………

**Date**……………………………

**CERTIFICATION**

We certify that this research work was carried out by **DINA ANJOLAOLUWA AFOLASHADE**

in the Department of Computer Science, College of Pure and Applied Sciences, Caleb University, Lagos. The research work is considered adequate in partial fulfilment of the requirements for the award of Bachelor of Science in Computer Science.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof. Moses Kehinde Aregbesola.**  Date

Project Supervisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Dr. Adeniyi Akanni**  Date

Head of Department

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof. Adetola Adesanya** Date

Dean of COPAS

**DEDICATION**

For God, who provided me with strength, grace and a whole lot more grace.

To my mother, who is the embodiment of what it means to never give up.

To my siblings and my Bug, whose support fuels my confidence.

And to myself, because I made it.

**ACKNOWLEDGEMENTS**

First, my most genuine, truest, deepest gratitude to Yahweh — for His grace, strength, mercies, and love. For being my Father, best friend, and guidance, all in one.

I want to thank my mother for being my unshaking foundation, and my siblings — Okikiola, for being here for me, unfailingly, every single day; Damilola, Titilayo, Gbemisola, and Olarenwaju — for all the different forms your support takes, for being my motivation and my pride. To my Bug, Mr. Gingerberry, thank you for your unwavering presence.

I cannot but acknowledge the current loves of my life — Aderonke Adesanya and Batifeori Omobolaji — who I couldn’t be happier to have call me Aunt.

Special thanks to my supervisor, Prof. M.K. Aregbesola, for your guidance and patience through this process.

To my beautiful friends — Jesica, Divine-Vision, Ameenat, Oluchi, Damilola, Seyi, Dami Talabi, Omoruyii, and all the guys — your presence made this lighter and easier.

And finally, back to God; because where else would I return?

**TABLE OF CONTENTS.**

DECLARATION……………………………………………………………………………….….i

CERTIFICATION ………………………………………………………………………………...ii

DEDICATION …………………………………………………………………………………...iii

ACKNOWLEDGEMENTS………………………………………………………...…………….iv

TABLE OF CONTENTS ……………………………………………………………....…………v

LIST OF FIGURES …………………………………………………………………………........x

ABSTRACT ……………………………………………………………………………………..xi

CHAPTER ONE: INTRODUCTION

* 1. BACKGROUND OF THE STUDY………………………………………………....1
     1. OVERVIEW OF AI BASED MULTIMODAL SYSTEMS……….…......2
     2. IMPORTANCE OF MULTIMODAL SENTIMENT ANALYSIS…….....3
  2. RESEARCH PROBLEM STATEMENT…………………………………….…..…..5
  3. RESEARCH AIM AND OBJECTIVES………………………….……………….....6
     1. SPECIFIC OBJECTIVES……………………….………………………..7
  4. RESEARCH QUESTIONS………………………………………….……………….7
     1. GENERAL RESEARCH QUESTIONS……………………………..........7
     2. SPECIFIC RESEARCH QUESTIONS……………………………….......7
  5. HYPOTHESES………………………………………………………………….……8
  6. SIGNIFICANCE OF THE STUDY…………………………………………..............8
  7. SCOPE AND DELIMITATIONS………………………………………………….....9
  8. THESIS STRUCTURE……………………………………………………………...11

CHAPTER TWO: LITERATURE REVIEW

2.1. INTRODUCTION……………………………………………………………...…...13

2.1.1. SIGNIFICANCE IN NLP, AFFECTIVE COMPUTING AND AI APPLICATIONS…………………………………………………………....…...13

2.2. SENTIMENT ANALYSIS (OVERVIEW)………………………....………….……14

2.3. HISTORY OF SENTIMENT ANALYSIS IN NATURAL LANGUAGE PROCESSING………………………………………………………………….………..15

2.3.1. EVOLUTION FROM UNIMODAL TO MULTIMODAL SENTIMENT ANALYSIS…………………………………………………………………… ...17

2.4. THEORETICAL FOUNDATIONS IN MSA…………………………………….....17

2.4.1. CORE ARCHITECTURES RELEVANT TO MULTIMODAL SENTIMENT ANALYSIS……………………………………………................18

2.4.2. MULTIMODAL FUSION STRATEGIES………………………………..21

2.4.3. OTHER THEORIES IN MULTIMODAL SENTIMENT ANALYSIS…..22

2.5. STATE-OF-THE-ART APPROACHES IN MULTIMODAL SENTIMENT ANALYSIS……………………………………………………………………………...24

2.5.1. RECENT ADAVANCEMENTS IN N.L.P FOR S.A……………………..24

2.5.2. RECENT ADVANCEMENTS IN COMPUTER VISION FOR S.A…….25

2.5.3. RECENT ADVANCEMENTS IN SPEECH PROCESSING FOR S.A….25

2.5.4. M.S.A: INTEGRATING TEXT, VISION AND SPEECH………………26

2.5.5. LEADING M.S.A. MODELS AND FRAMEWORKS………………….30

2.6. CHALLENGES IN M.S.A. AND SOLUTION APPROACHES………………….32

2.6.1. APPROACHES TO ADDRESS CHALLENGES IN M.S.A AND THEIR LIMITATIONS…………………………………………………………………35

2.7. GAP ANALYSIS IN M.S.A……………………………………………………….39

2.8. SUMMARY OF GAPS AND RESEARCH CONTRIBUTIONS…………………42

2.8.1. HOW DOES THIS HYBRID CONFIDENCE THRESHOLDING WITH LIGHTWEIGHT AMBIVALENCE FLAGGING APPROACH FILL THESE GAPS…………………………………………………………………………..43

CHAPTER THREE: METHODOLOGY

3.0. INTRODUCTION………………………………………………………………...45

3.1. MODEL SYSTEM ARCHITECTURE…………………………………………...45

3.2. RESEARCH DESIGN…………………………………………………………....47

3.3. DATA COLLECTION……………………………………………………………48

3.3.1. DATASET DESCRIPTION…………………………………………….48

3.3.2. JUSTIFICATION FOR DATASET CHOICE…………………………..50

3.4. DATA PRE-PROCESSING……………………………………………………….51

3.4.1 TEXTUAL DATA………………………………………………………..51

3.4.2. AUDIO DATA………………………………………………………….....51

3.4.3. VISUAL DATA…………………………………………………………....52

3.5. MODEL DEVELOPMENT…………………………………………………………52

3.5.1. CORE ARCHITECTURE AND TRAINING……………………………..52

3.5.2. PROPOSED MULTIMODAL MODEL (TEXT+AUDIO+VISUAL)……53

3.5.3. CONFIDENCE THRESHOLDING AND AMBIVALENCE …………....53

3.6. MODEL EVALUATION…………………………………………………………....54

3.6.1. CLASSIFICATION METRICS…………………………………………...54

3.6.2. MULTIMODAL EVALUATION METRICS……………………………..55

3.7. TOOLS AND FRAMEWORKS………………………………………………….....55

3.8. EXPERIMENTAL SETUP………………………………………………………….56

3.9. ETHICAL CONSIDERATIONS…………………………………………………....57

CHAPTER FOUR: RESULTS AND DISCUSSION…………………………………………....58

4.1. INTRODUCTION…………………………………………………………………..58

4.2. PRESENTATION OF RESULTS…………………………………………………...58

4.2.1. BASELINE PERFORMANCE…………………………………………...59

4.2.2. MULTIMODAL MODEL PERFORMANCE……………………………60

4.2.3. CONFIDENCE AND AMBIVALENCE RESULTS……………………...62

4.3. ANALYSIS OF RESULTS……………………………………………………......65

4.4. COMPARATIVE ANALYSIS WITH LITERATURE…………………………….66

4.4.1. PERFORMANCE COMPARISON WITH EXISTING WORKS ON THE MELD DATASET………………………………………………………………67

4.5. DISCUSSION OF FINDINGS……………………………………………………..70

4.6. IMPLICATIONS OF THE FINDINGS…………………………………………….72

4.7. CHALLENGES AND LIMITATIONS……………………………………………..73

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

5.1. SUMMARY OF THE STUDY……………………………………………………..75

5.2. CONTRIBUTIONS TO THE FIELD………………………………………………75

5.3. RECOMMENDATIONS FOR FUTURE RESEARCH……………………………77

5.4. PRACTICAL APPLICATIONS…………………………………………………….77

5.5. CONCLUSION……………………………………………………………………..78

REFERENCES…………………………………………………………………………………..79

APPENDICES…………………………………………………………………………………..88

**LIST OF FIGURES**

*Table 2.1. Summary of Benchmark datasets and transformer models in MSA……………………*31

*Figure 3.1. Model system architecture of the proposed model……………………………………...*.47

*Figure 3.2. Structure of distribution of MELD dataset………………………………………………*.50

*Figure 4.1. Precision, Recall, F1 score value table for baseline model……………………………*60

*Figure 4.2. Classification report with per class values for final MSA model……………………...*61

*Figure 4.3. Sample output showing high confidence prediction…………………………………….*63

*Figure 4.4. Low – confidence prediction with low confidence flag…………………………………*64

*Figure 4.5. Distribution of model confidence scores across test set predictions………………….*65

*Table 4.6. Comparison table with literature……………………………………………………………*69

**ABSTRACT**

*Multimodal Sentiment Analysis (MSA) presents as an essential advancement in affective computing. It addresses the limitations of the basic unimodal approaches that rely only on textual data. Traditional Sentiment Analysis approaches also struggle with ambiguity, emotional overlap and contextual nuance, which results in a reduction in accuracy when interpreting human emotion.*

*This research project proposes the design and development of a Multimodal Sentiment Analysis System that combines textual, audio and visual cues to improve the precision and robustness of sentiment detection. The system is trained and evaluated using the Multimodal Emotion Lines Dataset (MELD). This provided emotionally annotated dialogue clips that contain aligned transcripts, audio and video data. The proposed model architecture uses transformer based embeddings for text, acoustic feature extraction via OpenSMILE and facial emotion recognition using visual processing techniques.*

*An important innovation in this project is the adoption of the confidence thresholding and lightweight ambivalence flagging. These designs allow the addressing of emotional ambiguity and overlapping sentiment in multimodal data. Implemented using Python and Tensorflow, the system is evaluated based on standard classification metrics. These metrics include the accuracy, precision, recall and F1 score. This study also provides a framework for improving emotion-aware AI systems with practical implications for applications in various sectors not limited to healthcare, education and user-adaptive systems.*

*As this study focuses on both accuracy and emotional nuance, it contributes to the growing body of work seeking to make AI systems more emotionally intelligent, context-aware and grounded in ethics*

# **CHAPTER ONE: INTRODUCTION**

## **Background of the Study**

With every innovation in the digital world, it becomes more apparent that communication channels must support multiple modalities, including text, speech and visual expression. This is necessary to truly capture the richness of human interactions. This is because human sentiment is rarely limited to simply what is said, instead the full message is spread out across words, voice tone, pitch and other acoustic features and physical gestures (Poria et al., 2017). It has therefore been safely concluded that traditional sentiment analysis with its focus on textual input alone is inefficient. This insight has been the driving force of the evolution of sentiment analysis, moving from the previous text only analysis to systems that integrate and interpret emotional cues from multiple data sources.

The field of Multimodal Sentiment Analysis (MSA) intersects both affective computing and natural language processing. With the goal of improving accuracy and contextual understanding in sentiment prediction, MSA has combined textual data (what is said), audio data (how it’s said) and visual data ( including non-verbal cues). This helps to mimic how humans comprehend emotions (Zadeh et al., 2016). This approach to detecting sentiment is especially useful if systems are to be used in the real world, with improved applicability in areas like mental health support, customer feedback analysis, education technology and even social media monitoring.

Although a significant improvement from the unimodal sentiment analysis systems, the multimodal approach does come with its own set of challenges. These include emotional overlap, modality misalignment and ambiguous expressions that are rather difficult to classify as positive or negative without deeper reasoning (Morency et al., 2011). The field also remains relatively nascent, with ongoing opportunities to improve model accuracy and robustness. The most recent research has provided different ways to fuse modalities to tackle misalignments and improve accuracy as well as deep learning architectures to address some of these other issues. Still, there is usually a trade-off between accuracy and emotional nuance detection as simple models are often more accurate for basic sentiment tasks.

This project presents a multimodal sentiment analysis system that melds sentiment signals from text, audio and visual cues. As a broader goal, this work explores the integration of confidence thresholding and ambivalence flagging to help capture uncertainty and conflicting sentiment patterns. Grounding the system in existing multimodal benchmarks, addressing challenges of fusion and ambiguity and the use of these mechanisms helps the research contribute a flexible framework for sentiment analysis.

### **Overview of AI-based Multimodal Systems**

Artificial Intelligence (AI) has evolved significantly over the years, with simple rule based algorithms that were the norm, set aside for complex systems that are able to learn from vast and diverse datasets. This is done through a process known as machine learning. One of the major growth milestone in the field of AI is the processing of data of multiple modality including text, audio and visual data. The main purpose of this processing is to understand how humans are, and formulate better responses and reactions to their behaviors. The traditional AI systems worked with isolated single modality data with Natural Language Processing focused only on textual data, Computer Vision focused only on visual content and speech recognition focused only on audio inputs. It was obvious however that leaving the technology at that stage was not adequate, since human communication is inherently of a multimodal nature. Because of this, it became important that AI systems that can mimic this human-like understanding are developed (Baltrušaitis, Ahuja, & Morency, 2019).

Multimodal AI was birthed from this need. The distinguishing feature of these systems is that they meld information from various modalities and this in turn, allows for more content-awareness and emotional intelligence. This improvement allows AI systems to be more safely integrated in various valuable fields where a better understanding of a users emotional state has direct influence on the quality of the service being provided. Multimodal AI systems that integrate emotional intelligence bring us closer to creating more responsive and human-aware interfaces which is a crucial step in real world applications like virtual therapy, education, and emotionally sensitive services.

This advancement described above is made possible by the integration of some deep learning techniques. These techniques include the Convolutional Neural Networks (CNNs) for the processing of images, Transformers for language modelling and technology like the Recurrent Neural Networks (RNNs) and spectral methods for audio analysis. Using these techniques provides powerful tools for modelling complex relationship between modalities (Tsai et al., 2019). All of these are explored properly in this thesis paper, as well as other technology used in the multimodal SA world that has made it feasible to have such intelligence.

### **Importance of Multimodal Sentiment Analysis (MSA)**

MSA encompasses Natural Language Processing (NLP), computer vision and speech processing and each of these fields have found significant use in the development of intelligent systems. Isolated, these individual fields have made it possible for us to have machines that can understand text, interpret images and analyze audio signals. But as was established earlier, humans rarely communicate through a single channel. Because of MSA, our intelligent machines can extract, interpret and reason over emotional cues derived from multiple modalities in an organized integrated manner (Poria et al., 2017).

It can be summed up that the main importance of MSA is found in its potential to provide a more accurate and human-like understanding of affective states. Unimodal systems are liable to misunderstanding sarcasm, irony, complex or overlaid emotions or just generally contradictory signals and the MSA improves upon this by using the interplay between various modalities to reduce the effects of ambiguous data (Zadeh et al., 2017). For example, a sarcastic remark such as “Oh, I just love to be held back” can be interpreted textually as a positive statement but that may not be the case in audio analysis as the tone and pitch of words (flat) do not portray the excitement and positivity that the words do. It can be concluded therefore that classifying such a statement with a textual model alone will most likely misclassify it. If we were to take it a step-further and analyze the tone alongside the text, then the probability of misclassification reduces and it reduces further if we analyze facial expressions alongside text and audio. This puts MSA at a position to better grasp the true emotional intent behind the expression and thus communication in general.

MSA can be applied in almost every industry. The following examples provide insight on how MSA is implemented in different industries. In healthcare, one of the most important fields in the society, MSA can be used in mental health diagnostics (aiding in a sector of healthcare that is widely recognized as under-researched and underserved), patient monitoring and even in the utilization of therapeutic chatbots that are emotionally intelligent (Cummins et al., 2018).

In the field of education, MSA can be integrated in machines that power intelligent tutoring systems. These systems are able to detect when students are frustrated or confused and hence learn less and can then adjust teaching methods appropriately (D’Mello & Graesser, 2012).

MSA can also find extensive use in customer service, improving customer satisfaction by recognizing when they are distressed and responding in an emotionally aware and understanding manner. In security, we have behavioral monitoring systems that basically work by assessing the emotional changes of the involved parties in real-time and hence can serve as early warning tools for identifying threats or crises (Tzirakis et al., 2017). These can be used in therapy and interrogative sessions.

MSA also finds use in platforms like Amazon or applications that support streaming like Netflix, Spotify to provide customers with personalized recommendations.

One of the most evident importance, due to the rising popularity of virtual assistants and AI friends, is the use of MSA in human-AI interaction. Everyday, the interaction between humans and artificial intelligence takes a less formal, more conversational and emotionally nuanced path. Because of this, and the rise of social robots and metaverse platforms, the demand for emotionally intelligent systems grows. MSA provides a foundation for these systems, offering a strong and constantly improving basis for these systems and allowing them to be able to give personalized responses, improve engagement and ultimately build trust.

These diverse applications described above highlight exactly how valuable MSA is in real world systems, which serves as confirmation for the need for ongoing research that addresses the practical challenges of this field, including but not limited to resolving ambiguity, confidence based decision making and flexible multimodal integrations.

* 1. **Research Problem Statement**

In the pursuit to develop robots and systems that are more emotionally intelligent, it has been established that the unimodal sentiment analysis focused solely on a single modality is simply ineffective in classifying emotions that are nuanced, overlaid or complex in anyway. Constructs like irony, sarcasm, nostalgia, nervous excitement, bittersweet-ness etc. which are encompassed in emotional overlap and complexity of emotion, cannot be addressed or classified by unimodal systems and most often lead to the models defaulting to a dominant or more obvious cue which leads to inaccuracies in results that can have significant consequences when they are applied to real world scenarios or used in decision making.

While a more holistic approach is offered by the multimodal approach, the MSA systems in existence still face challenges that are related to accuracy, ambiguity, emotional overlap and uncertain predictions. The mechanisms for specifically handling conflicting emotions or overlap are rarely explored, leading to a drop in the degree of interpretability and robustness.

This research gap is addressed in this study by building a machine learning based multimodal sentiment analysis system that combines textual, audio and visual cues and integrates the confidence thresholding and ambivalence flagging mechanisms. The improvement of sentiment classification, while still accounting for emotional ambiguity is the ultimate goal here as this helps provide a system that is better suited to real world deployment.

* 1. **Research Aim and Objectives**

The main aim of this research work is to develop a multimodal sentiment analysis system that combines textual, audio and visual inputs while incorporating two important mechanisms; confidence thresholding and lightweight ambivalence flagging to improve robustness and emotional nuance recognition.

* + 1. **Specific Objectives**

1. Design a multimodal sentiment analysis system architecture that combines textual, audio and visual cues, integrating confidence thresholds and ambivalence flagging mechanism.
2. Implement the proposed multimodal sentiment analysis system using appropriate machine learning techniques.
3. Evaluate the systems accuracy and performance, assessing how effective the confidence thresholding and ambivalence flagging component are in improving the emotional intelligence of sentiment analysis systems.
   1. **Research Questions**

This study, aimed at the development of a multimodal sentiment analysis system that combines textual, audio, and visual cues is guided by the following research questions :

* + 1. **General Research Question**

How can a multimodal sentiment analysis system be designed to effectively combine textual, audio, and visual modalities for enhanced sentiment classification?

**1.4.2 Specific Research Questions**

a. What are the most suitable methods for collecting, processing, and aligning text, audio, and visual data for sentiment analysis?

b. How can information from these modalities be fused in a way that improves prediction consistency and interpretability?

c. Which deep learning architectures or components are best suited to handle the diversity of multimodal inputs?

d. How does the proposed system perform compared to traditional multimodal sentiment analysis models in terms of accuracy, precision, recall, and F1 score?

e. What are the core challenges encountered in integrating multimodal data, and how can these be addressed during system development and training?

f. In what ways can confidence thresholding be used to manage ambiguous or conflicting emotional cues?

g. How can ambivalence flagging enhance the model’s ability to detect and respond to emotionally uncertain inputs?

* 1. **Hypotheses**

H1: Multimodal Sentiment Analysis Systems that combine data from textual, audio and visual inputs will perform better than unimodal systems in detecting complex emotional states.

H2: The use of confidence thresholding and ambivalence flagging will improve the reliability of multimodal sentiment predictions in ambiguous emotional scenarios.

* 1. **Significance of the Study.**

This research holds significance in both the academic tables and practically, in industries. Academically, it adds knowledge to the growing body of work in the area of multimodal sentiment analysis and multimodal machine learning specifically, where it has become an important challenge to create models that understand what humans are thinking and feeling through multiple modalities (Poria et al., 2017). This study brings us closer to the broader goal of making AI systems that are better at understanding context and are simply more emotionally intelligent. It does this by proposing a machine learning based architecture that not only combines these modalities but also integrates ambiguity-handling mechanisms.

Introducing confidence thresholding and lightweight ambivalence flagging makes this study a distinct addition to current MSA architectures. These mechanisms improve system robustness by allowing for the detection and appropriate management of emotional overlap and ambiguity which are usually sources of error in the earlier models (Zadeh et al., 2018). By doing so, this research presents a more reliable, interpretable, and real world ready solution to affective computing.

From a practical standpoint, the system has wide relevance across industries increasingly integrating emotion aware AI. In healthcare, emotion-recognizing AI is now essential to mental health diagnostics and support tools that rely on accurate sentiment cues (Cummins et al., 2015). In education, emotionally adaptive systems respond to student frustration or disengagement in real time (D’Mello & Graesser, 2012). In customer service and marketing, sentiment-aware systems improve user experience by tailoring responses and recommendations to emotional context (Calefato et al., 2018).

By addressing the need for accuracy, emotional nuance, and robustness, this study helps bridge the gap between academic innovation and real-world deployment. It contributes to the development of next-generation HCI systems that are not only context-aware but also emotionally responsive. This is useful in helping the advancement of the field of emotionally intelligent AI in both theory and practice.

* 1. **Scope and Delimitations**

This project work is done with a focus on designing, developing, testing and evaluating a Multimodal Sentiment Analysis System that combines textual, audio and visual cues to improve the accuracy of emotion determination. Although there are several other modalities that can be integrated such as biometric data types like physiological signals ( such as heart rate and EEG) or even environmental context such as location and ambient noise, this study does not leverage any of these. The research scope is intentionally limited to textual, audio and visual data as stated above. This is because these modalities are the most accessible and commonly used signals in real-time, real-world Human-Computer interactions (Baltrušaitis, Ahuja, & Morency, 2019). The other modalities are excluded in a bid to keep things practical and implementation focused.

This research also explores confidence thresholding and lightweight ambivalence flagging a means of overcoming the problems of emotional overlap, complex emotional states and emotional confusion in sentiment analysis. And although these mechanisms offer innovative approaches for increasing the systems robustness, this work does not incorporate the full spectrum of techniques used in uncertainty modelling (such as Bayesian inference, ensemble modeling or Monte Carlo dropout methods) (Gal & Ghahramani, 2016).

Furthermore, this work uses already existing datasets rather than collecting raw data from the field. This has limited the extent to which the results can be generalized to other datasets or cultural contexts (as in accents and intonation), usually with the trade-off for efficient implementation.

Model evaluation will be carried out using pre-established metrics such as accuracy, weighted accuracy, precision, recall and F1 score. There is also emphasis on comparison between the model to be developed and early multimodal baselines as a means of evaluation. However, the design of this model is as a proof-of-concept prototype. It is not designed as a production-level system and thus may not be optimized for deployment constraints ( like robust error handling, security, logging and monitoring and deployment setup).

These delimitations are necessary to ensure a focused and achievable exploration within the time and resource constraints of this research.

* 1. **Thesis Structure**

This thesis is organized into five core chapters. A summary of each chapter is provided below:

A. Chapter One: Introduction

This chapter introduces the central research topic and presents the background of the study. It also provides an overview of artificial intelligence in the context of multimodal systems. Key elements covered include the research problem, objectives, research questions, significance of the study, scope and delimitations, and an outline of the thesis structure.

B. Chapter Two: Literature Review

Chapter Two provides a comprehensive review of existing work on unimodal and multimodal sentiment analysis. It examines core components such as natural language processing, speech emotion recognition, and facial expression analysis. Additionally, the chapter explores previous research on multimodal fusion strategies, deep learning architectures (including CNNs, RNNs, and Transformers), as well as known methods for managing emotional overlap, ambiguity, and complex emotional states. The chapter concludes with a gap analysis and a discussion of ethical considerations in MSA.

C. Chapter Three: Methodology

This chapter outlines the methodological approach used in the research. It covers the research design, data collection, data preprocessing, model development, model evaluation, tools and frameworks employed, experimental setup, and ethical considerations.

D. Chapter Four: Results and Discussion

This chapter presents the experimental results and provides in-depth analysis. It include sections on presentation of results, interpretation of system behavior, comparative analysis, discussion of findings, implications, and identified challenges or limitations.

E. Chapter Five: Conclusion and Recommendations

The final chapter offers a summary of the research findings and highlights the contributions of the study to the field of sentiment analysis and artificial intelligence. It also presents recommendations for future research and outlines practical applications of the proposed system.

**CHAPTER TWO: LITERATURE REVIEW**

**2.1. Introduction**

Sentiment Analysis is also known as opinion mining and it is a sub-field of Natural Language Processing that is concerned with the assessing and analysis of data; whether it be textual, visual or audio data. The main reason for carrying out this analysis is to identify and categorize sentiments, emotions or opinions. There are several techniques used in this process such as the lexicon-based methods, machine learning and the newer and most effective deep learning approach to determine if a given piece of writing is positive, negative, neutral or conveys a more complex, nuanced emotion (Liu, 2012). It has found essential use in the field of Artificial intelligence in the subfields of social media monitoring, customer feedback analysis and political opinion tracking.

**2.1.1**. **Significance in NLP, Affective Computing and AI applications.**

Sentiment Analysis (SA) plays a crucial role across Natural Language Processing (NLP), Affective Computing, and broader AI applications. Sentiment Analysis helps machines in NLP understand that the meanings of expressions vary with emotional context which in turn helps improve tasks like dialogue generation, translation and personalized content delivery. Furthermore, since the birth and integration of models like the BERT and GPT, the emotion detection capability of NLP systems have significantly improved as these models capture nuanced contextual information (Cambria et al., 2017; Devlin et al., 2019).

In the field of affective computing, SA serves as a foundational element seeing as affective computing is focused on building emotionally aware systems. It allows the development of applications like emotionally intelligent chatbots, virtual companions and tools, and mental health monitoring (Picard, 1997; Poria et al., 2017). The move toward MSA has improved the detection of complex emotional states even further.

Beyond these, SA is widely applied in AI-driven sectors including social media monitoring, market prediction, and healthcare. Through language cues, it allows brands to understand customer sentiment, guide financial forecasting, and supports mental health evaluation (Yadav & Vishwakarma, 2020). As these systems evolve, ethical issues such as bias and data privacy continue to be areas of concern and research (Hovy & Spruit, 2016).

Generally, SA remains central to building emotionally responsive AI systems. As research continues to address emotional ambiguity, cultural context, and multimodal complexity, its value will only deepen across real-world applications.

**2.2.** **Sentiment Analysis (Overview).**

Sentiment Analysis (SA), is also known as opinion mining. It is a subfield of Natural Language Processing (NLP) that detects, recognizes and classifies subjective information like emotions, attitudes or opinions by applying computational techniques. The expressions that are analyzed may come from textual, audio, or visual sources or a combination of modalities. As Liu (2012) explains, SA primarily seeks to classify sentiment as positive, negative, or neutral, and has gained popularity for its wide applicability across domains like business, healthcare, and human-computer interaction.

Traditionally, SA focused on textual data, using content that is user-generated like social media posts, reviews, or news to infer emotional tone (Pang & Lee, 2008). With time, audio-based SA was introduced and it pulled material for sentiment from tone, pitch, and speech patterns (Eyben et al., 2016). Visual-based SA is an even more recent addition, which analyzes facial expressions and body language, and is often used in combination with text and audio for greater emotional accuracy.

The most significant advancement in the SA space is the evolution toward Multimodal Sentiment Analysis (MSA). Unlike unimodal systems, MSA addresses challenges like sarcasm, ambiguity, and overlapping emotions by leveraging complementary data sources (Poria et al., 2017). This growth reflects the broader move toward emotionally intelligent AI systems capable of understanding human context more deeply.

**2.3. History of Sentiment Analysis in Natural Language Processing (NLP).**

Sentiment Analysis (SA) has undergone a remarkable transformation within Natural Language Processing (NLP). What began as simple rule-based systems has evolved into deep learning-powered multimodal frameworks that can detect emotion from textual, audio, and visual data.

1. **Early Stage: Rule-Based and Lexicon Methods (1990s–early 2000s)**

When sentiment analysis started, it relied on sentiment lexicons. These lexicons are a predefined list of words that are tagged as positive, negative or neutral. Turney (2002) tried to improve on this by using something called the Pointwise Mutual Information (PMI) to infer sentiment. This inference was done based on co-occurrence with seed words like bad or excellent. Soon after, machine learning based SA was introduced. Pang, Lee, and Vaithyanathan (2002) used algorithms like the Naïve Bayes, SVMs and Maximum Entropy models to classify sentiment in movie reviews. While some of the problems that came with the lexicon based approach was fixed with this, handling negation, sarcasm and domain specific context continued to pose problems.

1. **Shift to Machine Learning (Mid-2000s–2010s)**

More flexible models like the Decision trees and Logistic Regression were offered at the advent of supervised learning. Moving beyond binary classification, Pang and Lee (2005) were able to carry out scale based sentiment analysis.

As social media gained more momentum, Sentiment Analysis had to scale up. Research was conducted by people like Pak and Paroubek (2010) who used distant supervision and trained data labeled with emojis and hashtags. Still, online text contained a quality of informality and brevity that introduced noise and ambiguity.

1. **Deep Learning and Word Embeddings.**

Understanding context better through vector embeddings was made possible by the introduction of Word2Vec and GloVe (Mikolov et al., 2013). Soon after this, RNNs and CNNs were introduced. These technology outperformed the earlier models in detecting sentence level sentiment. Aspect-Based Sentiment Analysis (ABSA) also emerged, with a focus on sentiment toward specific features within a text (Pontiki et al., 2014).

1. **The Transformer Era (2017–Present)**

Sentiment Analysis as we knew it was completely revolutionized by the development of Transformers. BERT (Devlin et al., 2018) introduced bidirectional context processing, which improved the handling of negation and sarcasm. GPT models made SA even better by generating coherent, emotionally intelligent responses. Recently, these models are foundational in most cutting-edge SA tasks across fields.

1. **Multimodal Sentiment Analysis (MSA)**

As text alone does not truly convey emotional context, MSA combines text, audio and video, which offers a more complete picture of sentiment. Zadeh et al. (2017) introduced the Tensor Fusion Network (TFN) to model these modalities jointly. Newer works are exploring self supervised learning and multimodal transformers like MMBERT and CLIP. Emotion aware applications used in healthcare, education and human Computer Interaction are powered by these systems.

Simply put, from rigid lexicons to transformer-based multimodal models, sentiment analysis has evolved into a powerful tool for understanding affect. Today’s systems are not only more accurate but also more context-aware and emotionally intelligent even though they still face ongoing challenges like bias, ambiguity, and ethical concerns.

**2.3.1. Evolution from Unimodal to Multimodal Sentiment Analysis (MSA)**

The evolution from unimodal to multimodal that’s described in the history above was motivated by the fact that human emotion is not unimodal. And while the unimodal SA was a good start for a field that could be so much more, there was a need to beat the struggles with recognizing sarcasm, implicit sentiment, and ambiguous expressions that require more than just written words to decode. (Poria et al., 2017).

The developments described in the section above set the stage for the emergence of Multimodal Sentiment Analysis (MSA), which melds textual, audio, and visual data to capture the full spectrum of human emotion. Each modality plays a specific role, with the text capturing the literal message, audio adding vocal tone and prosody and visual cues like facial expressions and body language providing context. Combining these modalities improves robustness, contextual understanding and emotional nuance, thereby outperforming the unimodal systems in real world use (Mehrabian, 1971; Morency et al., 2011).

**2.4. Theoretical Foundations in MSA**

Multimodal Sentiment Analysis (MSA) is an interdisciplinary field. This research paper focuses on the computation and technological aspects but it would be remiss not to mention that MSA takes theoretical basis from some emotional theorems as well. MSA uses methodologies from several fields including Natural Language Processing (NLP), Computer Vision, audio processing and machine learning. These theoretical foundations come from several fields and they underpin how different modalities are processed to detect sentiment accurately.

**2.4.1. Core Architectures Relevant to Multimodal Sentiment Analysis (MSA)**

Several deep learning architectures are used in multimodal sentiment analysis to process and combine information from text, images and audio. This section provides information on the central architectures used in multimodal opinion mining. Some of these are the transformer-based models for textual sentiment analysis, CNNs and Vision transformers (ViTs) for identifying visual sentiment representations and acoustic feature extraction models for speech emotion recognition. Multimodal learning theories, datasets and fusion strategies have also been included.

* + 1. **Transformer Models for Textual Sentiment Analysis.**

Theoretical Basis: Transformer models take their theoretical basis from the self-attention mechanism. This mechanism is the reason why cross-modal is as effective as it is. It dynamically focuses on the most important features across the different modalities.

* + 1. BERT (Bidirectional Encoder Representations from Transformers)

BERT was developed by Devlin et al. (2019).. Its distinguishing bidirectional processing and self attention features make it such that BERT is in a better position to understand nuanced emotions that are expressed in text. (Sun et al., 2019). BERT as a model is usually trained on a very large amount of text data, and cam be fine-tunes for emotion categorization tasks. It uses datasets such as IMDb, SST-2, and SemEval (Hoang et al., 2022).

* + 1. GPT (Generative Pre-trained Transformer)

Another transformer model, this time developed by OpenAI, GPT is an autoregressive transformer model that generates human-like text (Radford et al., 2019). In its use in sentiment analysis, GPT has been fine-tuned on sentiment-labeled datasets to build on its capability to categorize feelings and “think up” sentiment -aware responses (Brown et al., 2020).

* + 1. Sentiment-Specific Transformer Variants

Seeing how effective the BERT and GPT are in the sentiment analysis field, more than one modification of transformers have been brought forward for sentiment analysis. ROBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) are examples of these proposed modified models and they optimize BERT’s architecture by adjusting hyper parameters and training strategies to enhance performance even more on sentiment categorization tasks.

* 1. **Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for Visual Sentiment Recognition.**

When it comes to visual sentiment recognition, BERT and GPT take a step back, allowing CNNs and ViTs to take the front seats. While the Vision Transformers (ViTs) take their theoretical foundation in self-attention mechanisms like other transformers do, the CNNs are completely different.

* + 1. Convolutional Neural Networks (CNNs)

Theoretical Basis: CNNs take their theoretical foundation from convolutional operations, like in the name. It uses these convolutional operations to identify spatial patterns and is therefore effective for extracting visual features and even audio spectrogram analysis.

In MSA, CNNs are responsible for the processing of facial expressions, body language, gestures and other visual cues to detect emotions (Fan et al., 2020). Certain models such as VGGNet developed by Simonyan & Zisserman, (2014) and ResNet (He et al., 2016) have been used in sentiment analysis tasks where they extract hierarchical characteristics from facial images and classify them into emotion classes such as happiness, sadness, anger and surprise.

* + 1. Vision Transformers (ViTs)

Instead of convolutional filters, ViTs use self-attention mechanisms (like in the text counterpart BERT), to capture long-range dependencies in images (Dosovitskiy et al., 2020). They have demonstrated superior efficiency in visual SA, especially when pre-trained on large-scale datasets like AffectNet and FER2013 (Kollias et al., 2021).

Hybrid CNN-ViT architectures (Khan et al., 2022) takes a best of both worlds approach; combining the pros of CNNs for low-level feature extraction and transformers for long-range dependencies, improving emotion categorization from facial expressions.

* 1. **Audio Feature Extraction Models for Speech Emotion Recognition**

Audio extraction works with audio data. Its main purpose is to identify emotions only from prosodic features. It is done using the following models:

* + 1. Wav2Vec

The speech emotion processing model, Wav2Vec, introduced by Baevski et al. (2020), is a self-supervised learning model. Its self-supervised nature takes its theoretical basis from Contrastive learning as well as representation learning It has been adapted since its innovation in 2020 for speech processing. This model has found application in speech recognition, where it detects variations in acoustic features such as tone, pitch, and prosody to categorize emotions in spoken language (Pepino et al., 2021).

* + 1. OpenSMILE

Theoretical basis: OpenSMILE uses prosodic and spectral feature extraction and statistical feature aggregation.

OpenSMILE, introduced by Eyben et al., (2010) is an handcrafted Audio Feature Extraction toolkit with the full meaning Open-Source Speech and Music Interpretation by Large-scale Extraction. It is vastly utilized as a feature extraction toolkit for speech emotion analysis. While deep learning models like Wav2Vec automate feature extraction, OpenSMILE remains valuable for traditional machine learning approaches in sentiment analysis (Trigeorgis et al., 2016).

**2.4.2. Multimodal Fusion Strategies.**

Theoretical Basis: Fusion theory takes its root from multi-view learning and information integration theory. This theory is of the opinion that merging a variety of data sources allows for robustness and accuracy of decision making processes.

There have been several fusion methods brought forward to make the modalities integration seamless, these are:

1. Early Fusion (Feature-Level Fusion)

In early fusion, the raw feature representations from the different modalities are combined before feeding them into a model (Baltrušaitis et al., 2019). That is, extracted text embeddings from BERT, image features from CNNs and audio features from Wav2Vec can be combined into a single feature vector before sentiment classification (Poria et al., 2015). Although there’s cross-modal interactions at an early stage, a main challenge with this strategy is that it requires precise synchronization of modalities.

1. Late Fusion (Decision-Level Fusion)

This strategy involves processing each modality independently and then combines the final predictions at decision level (Atrey et al., 2010). For instance, separate models may predict sentiment from text, images and audio and their predictions can be combined using majority voting or weighted averaging (Morency et al., 2011).. Computationally, this strategy is efficient, and even seems logical, but a real-life problem it has is catching cross-modal dependencies.

1. Hybrid Fusion (Intermediate-Level Fusion)

The Hybrid fusion takes the best of both worlds, balancing the strengths of early and late fusion by combining modalities at an intermediate stage (Poria et al., 2017). Multimodal transformers, such as the Multimodal Adaptation Gate (MAG) network (Rahman et al., 2020), dynamically adjust the influence of each modality based on contextual relevance. This strategy finds extensive use in addressing challenges like emotional Overlap and ambiguity.

**2.4.3. Other Theories in Multimodal Sentiment Analysis.**

There are several other theories integrated in MSA. Like was mentioned above, it is important to mention the psychological theories as emotion recognition deals with emotions and these theorems are crucial in the classification. The most important of these is the Ekman’s Basic Emotion Theory. This theory explains that the universal facial and vocal expressions correspond to emotions like happiness and anger which helps guide emotion classification.

Another such psychological theory is the Russell’s Circumplex Model of Affect. This one defines emotions along valence and arousal. That is, emotions are defined as positive or negative (valency) and have their intensify defined as well. It proposes that emotions are not objective, they come from an individuals interpretation of events. This helps shapes how textual, visual and audio cues should be analyzed together, and along with Ekman’s model, improves the emotional intelligence of MSA models.

MSA is also built on representation learning, cross alignment and sentiment classification theories. The first, representation learning theory makes sure that features are extracted from the various data sources effectively. Joint and coordinated representation learning lets us have models that can map modalities into a shared space while still keeping unique features.

The second, cross-modal attention and alignment theory works to solve synchronization issues by making sure that the modalities that are time sensitive such as facial expressions, are interpreted correctly using attention mechanisms.

Lastly, the sentiment classification and machine learning theory helps with categorizing across modalities more accurately using both contrastive learning and multitask learning.

Ethical considerations are also crucial in MSA as they are in every aspect of computer science. With MSAs, bias is very easy to be inherited by models based on the datasets they’re trained with. This is why when building proper MSAs, bias and fairness theories are implemented. They highlight the importance of addressing cultural and demographic variations in emotional expression to prevent biased predictions. Privacy-preserving machine learning is also utilized because ensuring the privacy of users is extremely important. It ensures ethical handling of multimodal data, particularly for sensitive applications involving video and audio.

**2.5. State-of-the-Art Approaches in Multimodal Sentiment Analysis.**

MSA has advanced a great deal since its onset with the integration of deep learning techniques in NLP, computer vision and speech processing. This section reviews the recent developments and state of the art approaches to the multimodal SA .

**2.5.1 Recent Advancements in Natural Language Processing (NLP) for Sentiment Analysis**

I. Transformer-Based NLP Models

Growth in the NLP based SA is largely brought about by transformer architecture and the self-attention mechanisms that they provide. They let us have models that outperform traditional machine learning models in understanding context, semantics and sentiment nuances .

1. BERT (Bidirectional Encoder Representations from Transformers)
2. GPT (Generative Pre-trained Transformer) models
3. Multimodal Transformers such as the Multimodal Transformer (MulT) which was debuted in (Tsai et al., 2019).

II. Aspect-Based Sentiment Analysis (ABSA).

Aspect-Based Sentiment Analysis (ABSA) refines traditional sentiment classification by identifying opinions tied to specific aspects of an entity (Xu et al., 2022). Transformer-based models such as T5 (Raffel et al., 2020) have been fine-tuned for ABSA tasks, allowing more granular sentiment analysis in user-generated content.

III. Sentiment Analysis in Low-Resource Languages.

Most sentiment analysis technology growth has been observed in the English market. Advancements in multilingual NLP while not as popular as in English, has led to increased performance in low resource languages. Models such as the mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) allow for cross-lingual emotion classification with limited labeled data.

**2.5.2. Recent Advancements in Computer Vision for Sentiment Analysis**

I. Vision Transformers (ViTs) for Emotion Recognition (Dosovitskiy et al., 2020).

II. AffectNet and Large-Scale Facial Emotion Datasets.

Deep learning models that can detect and identify very subtle nuances and fine-grained emotions in images are only brought into the realm of possibilities by the availability of large scale facial emotion datasets. Some of these datasets help in the training of models for peak performance. AffectNet (Mollahosseini et al., 2017) and FER-2013 (Goodfellow et al., 2013) are but a few.

III. Self-Supervised Learning in Visual Sentiment Analysis.

Recent studies in this field employs self-supervised learning (SSL) techniques in facial emotion recognition. SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) are able to pre-train models without the use of labeled data, greatly improving performance on small emotion recognition datasets.

**2.5.3. Recent Advancements in Speech Processing for Sentiment Analysis**

I. Self-Supervised Speech Models for Emotion Recognition

The speech-based sentiment analysis used to rely on handcrafted features extracted using toolkits like OpenSMILE in the past. With time, self-supervised models based deep learning models now perform way better than the traditional models. Wav2Vec 2.0 (Baevski et al., 2020) is another such model. HuBERT (Hsu et al., 2021) helps speech emotion detection by learning both phonetic and prosodic features. This makes it especially effective for identifying subtle emotional variations that’s present when humans speak.

II. Multimodal Speech Emotion Datasets

Nothing can be achieved in almost any field without datasets to base training on. For multimodal speech emotion, datasets like CMU-MOSEI (Zadeh et al., 2018) and MELD (Poria et al., 2019) have allowed researchers to innovate more robust MSA models by providing synchronized speech, text and video data. These data are then used in training models .

### **2.5.4. Multimodal Sentiment Analysis: Integrating Text, Vision, and Speech.**

Some noteworthy approaches here are:

* 1. Multimodal Transformer (MulT) (Tsai et al., 2019)

This transformer architecture learns cross-modal interactions through self-attention mechanisms, performing better than the traditional concatenation-based approaches. It was debuted in (Tsai et al., 2019) and helps align text, speech, and visual information for improved sentiment classification, demonstrating state-of-the-art performance on benchmark datasets like CMU-MOSEI.

* 1. CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021):

The CLIP is also a transformer architecture model. It contributes to the Multimodal Sentiment Analysis space by creating models that work by aligning image and text embeddings. CLIP is therefore great for synchrony of data forces. Because of these functionalities, CLIP has found use in recognizing emotions from memes and multimodal content.

* 1. FLAVA (Singh et al., 2022) :

FLAVA, Fusion of Language and Vision for Advanced Representation Learning is another transformer based model that works with visual and textual data sources. It builds on the capabilities of the CLIP by combining text (speech and captions) and visual modalities, improving performance in real-world sentiment analysis tasks. It helps tackle complexity in emotional states because it uses multimodal context when classifying emotions.

1. DialogueRNN (Majumder et al., 2019) :

This is the first RNN-based model that explicitly tracked speaker states in conversations. DialogueRNN uses GRUs and attention to model the intra- and interspeaker dependencies and while it was originally text-only, it’s extensions to multimodal input served as groundwork for later fusion techniques. The DialogueRNN uses GRU-based Party-State RNNs that are speaker-specific alongside Global-Context GRU (inter-speaker). It uses OpenSMILE, Long Short Term Memory and the FACET CNN. Its fusion mechanism uses attention over historical utterances.

1. MM-DFN (Liu et al., 2018);

The MM-DFN is a pioneer in efficient fusion. It employs low-rank factorization to reduce computational costs while dynamically weighting text, audio, and visual features. Its modest performance (WF1: 57.1%) highlights the trade-off between efficiency and complex multimodal interactions. MM-DFN uses BiLSTM (text) 1D CNN (audio/visual) as modality encoders. Its fusion is on low-rank modality specific factors along with dynamic attention. It contributed efficient fusion via parameter factorization.

1. ERHM (Hu et al., 2021) :

This model ( the SOTA hierarchical attention model ) integrates RoBERTa, Wav2Vec 2.0 and ResNet. All of these are leveraged using modality-specific feature extraction. Its performance demonstrates that there is value in combining advanced unimodal encoders with learned fusion strategies. The ERHM computes high accuracy, with its hierarchical attention (intra+inter model). This work contributed SOTA with advanced encoders.

1. Graph-MFN (Shen et al., 2021) :

The Graph-MFN is a hybrid of graph networks and memory fusion. It captures conversational structure through utterance level graphs. It outperforms pure RNN approaches as it bridges sequential and relational modeling. In this model, GCNs ( GloVe for text, VGGish for audio and ResNet for visual) are employed as the modality encoders. Fusion is carried out using graph memory propagation and the model is unique because of its hybrid graph and sequential modeling.

1. ICON (Hazarika et al., 2020):

This introduces interactive memory blocks that allow the modeling of long range multimodal dependencies in conversations. As a result of this innovation, ICON outperforms pure sequential approaches like the DialogueRNN while still maintaining interpretability. The value of explicit content storage for emotion recognition is highlighted by its gated memory mechanism. ICON uses the following modality encoders; BERT and DialogueRNN for text, OpenSMILE – LSTM for audio and FACET -CNN for the visual. It’s gated memory network stores cross-modal interactions and the interactive memory attention dynamically updates context.

Some other recent innovations in Multimodal Sentiment Analysis includes :

1. **Attention-Based Fusion Mechanisms**

Recent studies explore cross-attention fusion techniques to dynamically weigh the contribution of each modality based on context or the situation at hand (Rahman et al., 2020). For instance, a model may prioritize textual sentiment cues over facial expressions when analyzing sarcasm.

1. **Explainability in Multimodal Sentiment Analys**is

Explainability helps us break down and understand what the AI is focused on when it makes a certain decision and this can help long term to build trust, fix mistakes and ensure fairness. Knowing now the importance of explainability, it remains a challenge in deep learning-based MSA. Researchers have introduced attention visualization techniques (Das et al., 2020) and SHAP (Lundberg & Lee, 2017) to interpret which modalities contribute most to sentiment predictions in an attempt that the question of why, can be answered.

**2.5.5. Leading MSA Models and Frameworks.**

|  |  |  |
| --- | --- | --- |
| DATASETS /MODEL | CORE CONTRIBUTIONS | LIMITATIONS. |
| CMU- MOSEI Dataset. | 1. Provides a large dataset with fine -grained emotions annotations ( Zadeh et al., 2018). 2. Allows sentiment analysis at both the utterance and video levels. 3. Complex emotional expressions can be modeled due to the vast bank of diverse linguistic, visual and acoustic data it provides. | 1. Models may have problems handling noisy data as a result of real-world variations in speech and facial expressions 2. The dataset also contains subjectivity in emotion annotations which leads to ambiguities in model training. |
| M3ER (Multi-modal Multi-label Emotion Recognition) | 1. Adaptive fusion strategy is used to compare the contributions of the different modalities dynamically (Mittal et al., 2020). 2. addresses emotional overlap by using a multi-label categorization approach ( Mittal et al., 2020). | 1. Deep neutral networks and models based on them like the M3ER require a large amount of computational resources. 2. Faces challenges synchronization errors when melding multimodal data from different time frames. |
| Multimodal Transformers (MMT, MISA, MAG-BERT, etc.) | 1. Self-attention mechanisms improve feature alignment between modalities. 2. enhance generalization on unseen data. | 1. High computational complexity limits real-time applications. 2. Require large-scale labeled datasets to achieve optimal performance. |

Table 2.1: Summary of benchmark datasets and transformer-based models used in Multimodal Sentiment Analysis (MSA), including their core contributions and key limitations. The table highlights how each dataset or model advances sentiment classification while also showing their implementation constraints.

**2.6. Challenges in Multimodal Sentiment Analysis and Solution Approaches.**

Despite the advancements in the multimodal Sentiment Analysis technology, several challenges remain that influence its effectiveness. These challenges include emotional overlap, complex emotional states, modality fusion issues, data annotation difficulties and ethical concerns.

i. Emotional Overlap and Ambiguity

Emotions overlap describes a situation where we as humans feel more than one emotion at a time. Realistically, there is an overlap of emotion in almost every moment of life and it shows in the expressions and audios that are to be analyzed. One of the main challenges in MSA is overcoming this. When multiple emotions coexist within a single expression, it becomes difficult for models to give a definitive sentiment label (Akhtar et al., 2019).

It is easier in unimodal sentiment analysis because the data is from a single modality and there are no dependencies. This is obviously not the case in MSA and so MSA must find ways to handle situations where facial gestures convey a certain sentiment while their tone of voice suggests another (Poria et al., 2020). This problem makes model training and evaluation difficult and hence leads to sentiment predictions that are inconsistent.

Another term thrown in with overlap in MSA is ambiguity. It can be said that emotional Overlap leads to ambiguity and in turn impacts the models. Someone might say, “I’m fine” with a mix of relief and frustration, which contains both traditionally positive and negative sentiment, making it challenging to determine their true emotional state without additional context. Emotional Overlap can not be ignored because it plays a big role in how accurate the model will be. If a model fails to identify overlapping emotions, it may classify it wrongly or oversimplify it into a single category.

ii. Complex Emotional States

Another challenge that is presented in Multimodal Sentiment Analysis is the problem of complex emotional states. This phenomenon is somewhat similar to overlap and ambiguity but there are still significant differences. If MSAs are to be integrated successfully in everyday life, models that can identify emotions that go beyond simple positive, negative or neutral sentiments must be developed. The reason behind this being that real world sentiment analysis involves determining complex emotions such as sarcasm, irony, nostalgia and mixed feelings. These emotions are especially difficult to recognize and classify using traditional SA models (Zadeh et al., 2018). For example, sarcasm usually involves a mismatch in the spoken words and in the speaker’s tone or facial expressions, and Multi-modal sentiment analysis systems must accurately interpret to avoid wrongly classifying (Hazarika et al., 2020).

1. Modality Fusion and Alignment Issues

Another impossible to ignore challenge in MSA is the integration and alignment of different modalities. Each modality has different temporal and spatial characteristics and this makes it difficult to synchronize them effectively (Baltrušaitis et al., 2019). For example, speech occurs continuously over time, while textual data is discrete and facial gestures can change dynamically. A misalignment in these modalities can cause a loss of information or inconsistencies, decreasing the model’s accuracy (Sun et al., 2020).

1. Data Annotation and Labeling Complexity

Data annotation refers to the process of labelling data to train machine learning models. Labelling complexity is a term that describes the challenges involved in accurately and effectively annotating data. These challenges may arise as a result of various factors such as ambiguity in data (a sentence like “I’m fine” having multiple valid labels ranging from neutrality to sarcasm to mixed emotions) (Pérez-Rosas et al., 2013). In MSA, emotional cues from text, audio and visual data may clash this making annotation difficult. Another factor contributing to Labelling complexity is subjectivity of labels as a result of cultural, personal or contextual biases. It has been noted that inter-annotator agreement is often low for subjective tasks like emotion recognition (Soleymani et al., 2017).

Multi-label and single-label classifications may also influence labelling complexity as some complex situations will require multi-label annotation. An example is a speech clip expressing both happiness and nervousness. Granularity of labels with broad category labels such as happy, sad and angry falling short in terms of representing nuanced emotions such as bittersweet, melancholic and relieved, is also a set back to data annotation. Other setbacks are cost and scalability as manual annotation of large data is expensive and time consuming and automated labelling introduces errors.

1. Ethical Concerns and Bias in MSA

It is a general knowledge that models inherit bias from datasets they are trained on. This therefore introduces ethical concerns in MSA on the topic of biases in training data(which may be due to subjectivity in labelling as discussed above), privacy issues and the potential abuse of Sentiment Analysis models. Sentiment classification models trained on biased datasets may strengthen already existing prejudices, leading to unfair or discriminatory outcomes (Mehrabi et al., 2021).

There are also concerns regarding user consent and privacy when analyzing sentiments from personal conversations, facial expressions or voice recordings without explicit permission (Crawford & Calo, 2016).

In essence, MSA is great for improving accuracy and greater applicability but every advancement comes with a set of challenges. Addressing these challenges is essential for improving the reliability, fairness and accuracy of these models. Future research should put emphasis on developing more diverse fusion techniques, models that better handle ambiguity and complex emotional states, standardizing annotation methodologies and implementing ethical guidelines to mitigate bias and privacy concerns.

**2.6.1. Approaches to Address Challenges in Multimodal Sentiment Analysis and their Limitations**

Already, many researchers have turned their attentions toward trying to solve the challenges that plague the MSA field. Their research and efforts have yielded several proposed solutions to tackle the challenges in MSA. It is important to note though, that these solutions are partial solutions as they improve sentiment predictions but come with limitations that influence their generalizability, interpretability and ethical considerations. Below are some of the key approaches and their limitations:

**i. Handling Emotional Overlap and Ambiguity.**

To handle the problem of emotional overlap and ambiguity, researchers have proposed the following:

a. Approach: Confidence Thresholding and ambivalence flagging.

Confidence threshold and ambivalence flagging have been proposed to help solve the problem of emotional overlap and ambiguity. Some studies, favor the confidence thresholding where emotion predictions below a certain confidence level is flagged as ambiguous (Poria et al., 2020). This makes the system acknowledge ambiguity instead of forcing inaccurate categorization.

Limitations: Confidence thresholds are arbitrary, not following a laid down system and may even be considered whimsical based on the researcher, based on personal choice. They are also often data-set dependent, making them difficult to generalize across different datasets. They also do not give a deep understanding of emotional overlap, only filtering uncertain cases.

b. Approach: Multi-Label Classification Models

Another approach developed to address the emotional overlap challenge is the Multi-label classification model. Researchers have explored multi-label classifiers that allow assigning more than one emotion to a single instance (Akhtar et al., 2019).

Limitations: The Multi-label approach requires very vast, very detailed, emotion-rich datasets and such datasets are hard to come by, given the sheer amount of resources that would be required to annotate them. Multi-label classification models also increases computational complexity, making it difficult to present such approaches to real world applications.

**ii.** **Detecting Complex Emotional States (Sarcasm, Irony, Mixed Emotions)**

The problem posed by complexity of emotions is currently being battled thus:

Approach: Context-Aware Transformers (BERT, RoBERTa, MISA)

Recently, studies have delved into deep learning models like MISA and transformer models like BERT and RoBERTa to adequately represent contextual nuances (Hazarika et al., 2020). These models leverage self-attention mechanisms as discussed above to identify more complex emotions like sarcasm, irony, nostalgia, melancholy etc..

Limitations: Models based on transformer technology require extensive training data, which may not be available for low-resource languages. These models also require high computation and its use may not be plausible for real -time applications. They often act as “black-box” models, making it difficult to explain predictions.

**iii. Modality Fusion and Alignment Issues**

These are issues that arise when researchers try to answer the question “how exactly do we combine modalities?”

a. Approach: Early, Late, and Hybrid Fusion Techniques

There are three fusion strategies currently in use in the MSA field; early fusion, late fusion and hybrid fusion.

Limitations: Early fusion struggles with feature incompatibility, as different modalities have varying structures and temporal resolutions. Late fusion may not fully capture cross-modal interactions, decreasing the synergy between text, audio, and video. Hybrid fusion does improves performance but it requires complex architectures that are difficult to optimize.

b. Approach: Cross-Modal Attention Mechanisms

Cross-modal transformers and attention-based mechanisms allow models to focus on the most relevant aspects of each modality, improving alignment (Sun et al., 2020).

Limitations: This approach is heavily dependent on high-quality annotations, making training expensive and time consuming. It is also computationally intensive, making real-time deployment challenging.

1. **Data Annotation and Labeling Complexity**

Several approaches have been developed in an effort to reduce the sheer amount of resources that go into data annotation.

a. Approach: Weak Supervision and Semi-Supervised Learning

In an effort to reduce the high costs of manual annotation, weak supervision techniques that leverage heuristic rules, distant supervision or pre-trained models for labelling were introduced (Pérez-Rosas et al., 2013).

Limitations: The labels generated using this approach are noisy, reducing the general reliability of training data. It is often biased toward the heuristics used for label generation, leading to systematic bugs.

b. Approach: Crowdsourced Annotation & Active Learning

Crowdsourcing platforms like Amazon Mechanical Turk allow large-scale data labeling (Soleymani et al., 2017). Active learning selectively queries annotators for the most uncertain samples.

Limitations: Crowdsourced labels may be inconsistent due to varying annotator expertise. Active learning still requires human intervention, limiting full automation.

1. **Ethical Concerns and Bias in MSA**

As described in the previous section are issues that arise as a question of the integrity of MSA systems. Some noteworthy approaches are:

a. Approach: Bias Mitigation via Fair Representation Learning

Techniques such as adversarial learning aim to decrease bias by ensuring fair representations across demographic groups (Mehrabi et al., 2021).

Limitations: There is a very real chance that this approach may decrease the accuracy of models in an attempt to remove bias. Also, bias in the pre-trained embeddings can still propagate irrespective of the mitigation efforts.

b. Approach: Privacy-Preserving Sentiment Analysis

Federated learning and differential privacy techniques allow models to learn from decentralized user data without direct data exposure (Crawford & Calo, 2016).

Limitations: The federated learning requires substantial coordination and infrastructure. Differential privacy techniques can also introduce noise, reducing model performance.

Simply put, each approach to solving MSA challenges comes with a set of pros and cons. Future research should focus on developing more interpretable, resource-efficient models that can handle complex emotional states, improve alignment and reduce bias while still keeping ethical integrity.

**2.7. Gap Analysis in Multimodal Sentiment Analysis.**

Even with all the technological and research innovations in Multimodal Sentiment Analysis, there are still gaps in research, especially in the handling of emotional ambiguity, leveraging confidence thresholding mechanisms and adapting models to real-world conditions. Improving the performance accuracy, interpretability, robustness, generalization and real-life applicability of MSA systems is based on researchers ability to address these gaps.

1. Limited Handling of Emotional Overlap

One of the main problems in MSA is the difficulty that models have in accurately classifying overlapping emotions. Human emotions are most of the time very complex, with multiple affective states occurring at the same instance (Akhtar et al., 2019).

The original sentiment analysis methods worked using single -label categorization, which basically makes it such that a model has to assign only one sentiment per instance, disregarding others, even when multiple emotions coexist (Poria et al., 2020).

Even though multi-label classification has been explored, it remains limited due to:

1. The lack of proper resources to produce high quality annotated datasets that capture emotional overlap and nuances in emotion.
2. Interdependences that may exist between emotions such as how “love” and “anger” may coexist in a speaker’s tone. There is difficulty in modelling such interdependencies.
3. It has also been established that there is a tendency for models to favor dominant emotions, which means that the more subtle emotional cues are ignored (Soleymani et al., 2017).
4. Lack of Confidence Thresholding for Ambiguous Cases

A gap in MSA field is the lack of confidence thresholding mechanisms to handle ambiguity or uncertainty. Sentiment models typically generate a prediction with no indication of just how sure they are of that classification(Hazarika et al., 2020). The issue in this lays in the fact that emotions exist on a spectrum and certain instances may not fit neatly into predefined sentiment categories.

Existing approaches attempt to improve model confidence through ensemble methods or attention mechanisms (Zadeh et al., 2018), but they do not plainly tell when a model should abstain from making a decision. Confidence thresholding would allow models to flag ambiguous cases so that they can be further examined, which improves overall accuracy.

Some researchers have explored uncertainty estimation techniques, such as Bayesian neural networks (Sun et al., 2020), but these models are not widely used because in terms of computational complexity, they are quite very expensive. Moreover, there are no standardized approach for integrating ambivalence flagging in MSA frameworks.

1. Real-World Data Constraints and Generalization Issues

Many MSA models are developed, trained and tested using controlled, benchmark datasets, such as CMU-MOSEI and IEMOCAP, which contain well-annotated multimodal data (Pérez-Rosas et al., 2013). It is important to note though, that these datasets as robust as they may present do not manage to encapsulate the unpredictability and messy-nature of real-life day-to-day interactions where things such as mentioned below are present.

* 1. Noisy data (such as background noise in audio, occluded faces in video) affects sentiment analysis and detection.
  2. Cultural and linguistic differences impact how emotions are expressed.
  3. Spontaneous human interactions differ significantly from scripted or staged datasets.

Models trained on clean, well-segmented datasets therefore often fail to perform very well when applied to real-world, in-the-wild settings (Baltrušaitis et al., 2019). Presently there are only a handful of studies that explore domain adaptation techniques to bridge this gap.

**2.8. Summary of Gaps and Research Contributions.**

In short, after a careful review of the existing works in MSA, it can be summarized that many systems struggle to differentiate between similar emotions like frustration and sadness as a result of overlapping features in text, audio and visual cues. Many systems also fail to detect mixed or ambivalent emotional states, leading to inaccurate predictions or an incomplete sentiment classification.

MSA also faces challenges in confidence and uncertainty handling, as most models predict sentiment without indicating their confidence level and this can result in unreliable classifications especially in the case of ambiguity. The current models often wrongly classify low-confidence inputs instead of acknowledging uncertainty, impacting system reliability. It is evident that the distinct combination of confidence thresholding with lightweight ambivalence flagging in MSA is not extensively explored. While various studies have proposed advanced fusion strategies and deep learning frameworks to enhance SA accuracy, the combination of a confidence threshold mechanism, alongside a targeted efficient ambivalence detection approach remains relatively novel.

And while there has been attempts to fill the ambiguity gap, detecting ambivalence usually requires complex models that are computationally intensive, making them impractical for real time application. Research in lightweight methods to handle emotional overlap remains limited. For example, a study titled “Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis” introduces a novel fusion strategy to improve audio-text SA accuracy. However, it does not specifically address the leveraging of confidence thresholds or ambivalence flagging. In a similar manner, the survey “Multimodal Sentiment Analysis: A Survey” has provisioned an overview of the growth of the MSA, covering recent datasets and advanced models, but does not dive into this specific hybrid approach that is developed in this research.

Finally, misclassification in sentiment analysis, especially in sensitive areas like mental health or customer service, can have significant negative impacts, yet many systems do not leverage mechanisms to minimize harm in ambiguous situations.

Therefore, this research on implementing a hybrid system that combines confidence thresholding with lightweight ambivalence flagging fills a notable gap in the current MSA literature, offering a unique contribution to the field.

**2.8.1. How Does This Hybrid Confidence Thresholding with Lightweight Ambivalence Flagging Approach Fill These Gaps:**

This research work builds a Multi-modal sentiment analysis system that combines textual audio and visual cues and uses confidence thresholding and lightweight ambivalence flagging to overcome overlap and emotional complexity. It fills gaps in existing literature thus:

1. Addressing Emotional Overlap and Ambiguity:

This study introduces a lightweight ambivalence flagging system. Lightweight, in order to not be hindered by the size of the computational power required for a full size ambivalence flagging system while still tackling the issue of the emotional overlap head on. Instead of forcing systems to choose one emotion when multiple are present, this model flags cases where emotions are too close to distinguish confidently, thereby improving accuracy in complex emotional states.

1. Handling Uncertainty through Confidence Thresholding:

This model also leverages a confidence thresholding system that makes sure that only predictions above a certain reliability level are accepted. Ambiguous inputs below the threshold are flagged, reducing the likelihood of misclassification and making the system more transparent and reliable.

1. Enhancing Ethical Reliability in MSA:

By flat-out flagging uncertain and ambiguous cases, this model helps mitigate ethical risks associated with misclassifications. This adds to the growing need for responsible AI systems, especially in applications where SA can impact the well-being of the user such as suicide chatbots or business decisions.

Simply put, this approach directly addresses gaps in emotional ambiguity, reliability, computational efficiency, and ethical handling in MSA. It proposes a practical and innovative solution by combining a confidence threshold system with a lightweight ambivalence detection layer, making sentiment analysis more accurate, transparent, and ethically sound.

**CHAPTER THREE: METHODOLOGY**

3.0. **Introduction**

In this chapter, the research methodology employed in the design, implementation and evaluation of a Multimodal Sentiment Analysis System that combines textual, audio and visual cues is presented. The information included in this chapter is aimed at providing a detailed account of how the system was conceptualized, actually implemented and tested. Below, the research design, datasets used, data pre-processing approaches, and details on the model architecture used for each modality are explained. Furthermore, sections in the chapter explore the design and the model architecture for the integrated model.

This chapter also covers discussions on the tools and frameworks used, elaborating on the experimental set-up. The evaluation metrics that are to be used to access the systems performance are given, and it highlights relevant ethical considerations.

This section forms the backbone of the research. It enables the development of a robust, context aware and emotionally intelligent sentiment analysis system, making it more applicable to real life.

**3.1. Model System Architecture**

The proposed Multimodal Sentiment Analysis model is designed to reflect how humans perceive emotions. Its architecture is broken down into five key modules. These different components are then fused into a single pipeline to make sure that there is synchrony in analysis of textual, acoustic and visual cues for each utterance.

1. Text Processing Module: The BERT tokenizer (bert-base-uncased) is used here, to tokenize each of the utterances. It generates a 768-dimensional embedding vector that represents the semantic content.
2. Audio Processing Module: Using moviepy, audio is extracted from the original video and then relevant acoustic features, the MFCCs, are extracted using the Librosa library. In situations of missing audio, smart imputation methods are adopted. This includes imputation based on similar emotions or nearest neighbors.
3. Video Processing Module: Every video clip is broken down into 10 evenly spaced frames. Visual features are then extracted from each frame using a pretrained ResNet50 model. These are mean-pooled to produce a 2048-dimensional vector thereby capturing the overall visual sentiment.
4. Fusion and Classification Module: The vectors that are outputted by the three modalities are concatenated to form a single, unified multimodal feature vector. This means that fusion is early fusion by concatenation. This vector is then passed through a classification network. The final classifier is a custom MLP in Pytorch. This network outputs predicted emotion and sentiment classes.
5. Post-Processing Module: Predictions that fall under a certain certainty margin are flagged using the confidence thresholding mechanism. The ambivalence flagging goes further to detect overlap or ambiguity in emotional signals. Incorporating these allows the model account for emotional complexity, thereby providing a more transparent, real-world- applicable result.

*Table 3.1. Model System Architecture of the proposed model.*

**3.2. Research Design**

This research work takes a model-based experimental and analytical approach. This approach is focused on developing an MSA system that combines textual, audio and visual cues with an aim to improve the accuracy and emotional depth of sentiment categorizing. Furthermore, the system goes beyond emotion detection only, integrating confidence thresholding and lightweight ambivalence flagging in a bid to address the problem of emotional ambiguity and overlap.

The proposed multimodal model is compared to the unimodal baselines (text-only, audio-only, visual-only) in this experiment. This comparison is carried out based on standard performance metrics such as the accuracy, precision, recall and F1 score. Including the confidence thresholding mechanism enables the model to flag down outputs in uncertain or conflicting scenarios. The ambivalence flagging then highlights emotional signals that are mixed or overlapping for further interpretation. These additions are in a bid to increase real world applicability by improving how well the system handles complex emotional expressions.

This research project is implemented using the Python programming language and is built on deep learning architectures. Frameworks like Hugging Face transformers and Pytorch also allows diverse data modalities to be integrated into a single, unified model pipeline. These frameworks also support the implementation of threshold based logic and flagging mechanisms quite flexibly.

This experimental setup is justified by the fact that human emotions is inherently of a multimodal and ambiguous nature. Thus systems that can understand, adapt and sometimes even admit uncertainty in interpretation are needed (Morency et al., 2011; Zadeh et al., 2017). By integrating these mechanisms for confidence estimation and ambiguity detection, this study adds progress to the development of emotionally intelligent AI that is more reflective of how communications happen in the real world.

**3.3. Data Collection**

The dataset employed in this study is the Multimodal Emotion Lines Dataset (MELD). It is the sole dataset used for the training, validation and evaluation of the proposed multimodal sentiment analysis system. MELD is a benchmark dataset that is curated specifically for emotion recognition in multiparty conversations. It is therefore well suited for works that involve multimodal sentiment detection using textual, audio and visual data.

**3.3.1. Dataset Description**

MELD was introduced by Poria et al. (2019). It is derived from the popular TV series called Friends. The dataset consists of over 13,000 utterances pulled from 1,433 dialogues between multiple speakers. Each of the utterances is annotated with emotion and sentiment labels that are aligned with text transcripts, corresponding audio recordings (WAV) and video clips (MP4). Speaker identifiers as well as dialogue level context are also contained in MELD, enabling us model not only isolated emotions but also dynamic emotional interactions in conversations.

The emotion label categories include Neutral, Joy, Anger, Sadness, Disgust, Fear and Surprise. The sentiment label categories includes Positive, Negative and Neutral. The data is split into three sunsets thus:

1. Training set (9,989): This is the data used in the training of the model. It is for learning and optimization.
2. Validation set (1,109): The data in this set is for fine-tuning after training. It is also aimed at helping to avoid overfitting.
3. Test set (2,610): The data in the test set is used for evaluating model generalization on unseen data.

MELD allows models to analyze the interplay between verbal content, vocal prosody and facial cues in emotional communication because of its multimodal structure. This fits directly with the directions of this research, which is to improve sentiment recognition accuracy and robustness of models, by combining all three modalities.



*Figure 3.2 Screenshot showing the structure of the distribution of MELD dataset samples across training, validation, and test splits.*

**3.3.2. Justification for Dataset Choice.**

MELD is chosen due to its comprehensive multimodal alignment. It also offers context rich dialogues and emotional diversity, all of which make it very suitable for capturing complexity and overlapping of emotions in expressions. Since this research’s proposed system makes use of mechanisms like the confidence thresholding and ambivalence flagging, MELD provides a sufficiently rich structure to test and validate these components.

MELD is also available publicly and is widely applied as a benchmark in MSA research as it supports comparability and reproducibility. Utterance level granularity as well as conversation level coherence are supported by MELD, which allows the system an opportunity to learn not only isolated emotional states but also the transitions and ambiguities that are present in real life interactions.

**3.4. Data Pre**-**processing**

The MELD dataset underwent modality specific preprocessing steps for text, audio and visual data to ensure that there is effective training and evaluating of the proposed model. Since MELD provides aligned transcripts, audio, and video, pre-processing is carried out to optimize the input quality and to synchronize features across the three modalities.

**3.4.1. Textual Data**

Text pre-processing was carried out using standard Natural Language Processing techniques. The specifics are detailed below:

1. Tokenization: This process involved the use of bert-base-uncased from Hugging Face Transformers. Each utterance produces a 768-dimensional vector. Embedding Generation involved the inputted texts converting into dense vector representations. Embedding was carried out by mean pooling from last\_hidden\_state after truncating to 128 tokens. The conversion, being done using pre-trained transformer model BERT, is capable of capturing the semantic and contextual relationships between different words (Devlin et al., 2019).

**3.4.2. Audio Data**

The audio component was preprocessed to extract emotional acoustic features.

1. Feature Extraction: Moviepy was used to extract “.wav” files from MP4 files. The library Librosa was used then used to extract an important acoustic feature. This feature is the Mel-Frequency Cepstral Coefficients (MFCCs) with 13-dim. Emotion-based or KNN imputation method was used to fill gaps in feature vectors in cases of missing audio files. This pre-processing activity outputs a fixed-size audio feature vector per utterance. These features help represent tone, stress, and rhythm, which are often correlated with emotional cues.

**3.4.3. Visual Data**

The visual inputs were cleaned thus:

1. Frame Extraction and facial emotion recognition: Ten frames per clip were first extracted in a uniform manner. Then, features were extracted using pretrained ResNet50. This was then mean-pooled into a 2048-dimensional vector.

**3.5. Model Development**

This section gives an outline of the baseline, hybrid and proposed multimodal models developed for sentiment analysis using textual, audio and visual inputs. It is aimed at evaluating how well each modality performs independently and in combination. It’s also set to help determine whether combining modalities improves emotion detection accuracy and robustness.

**3.5.1 Core Architecture and Training.**

Unimodal models were developed for each modality in order to establish a performance benchmark.

1. Text-only Model:

For processing textual data, a BERT-based sentiment classifier (bert-base-uncased) was implemented. The model uses contextual embeddings that are generated from a pre-trained BERT model that are then fine-tuned and passed into a dense classification layer. Each one of the utterances is represented as a 768-dimensional embedding. This helps with modularization and serves as the baseline for understanding the contribution of text to the overall sentiment prediction.

1. Audio-only Model:

For the audio-only setup, acoustic feature Mel-Frequency Cepstral Coefficients (MFCCs), were extracted from the speech data using Librosa. After extraction, they are used to learn the patterns in vocal tone and prosody that are indicative of emotional states.

1. Visual-only Model:

In the visual-only approach, image frames were extracted from video clips and processed using a pretrained ResNet50 to detect facial expressions and other visual cues. Each video is then represented by 2048-dimensional vectors derived from mean pooling across selected frames. Non- verbal emotions like smiles, frowns and raised eyebrows, which are essential for identifying sentiment are all captured in this way.

**3.5.2 Proposed Multimodal Model (Text + Audio + Visual)**

The final proposed model integrates all three modalities. It carries out early fusion by concatenating features from all the three modalities (text, audio, visual). This unified representation is then passed into a final classifier which is a custom Multi-Layer Perceptron (MLP) in Pytorch with 5 linear layers + dropout + ReLU + batchnorm, with hidden layers and a softmax output layer. The model is then trained using cross-entropy loss and optimized using Adam optimizer (lr=1e-4), 15 epochs, loss is Weighted CrossEntropyLoss, scheduler is StepLR. All of this is implemented through the Pytorch framework in Python.

**3.5.3 Confidence Thresholding and Ambivalence Flagging.**

Upon prediction generation through the softmax later, a confidence score is assigned to each class. This mechanism was integrated to help identify predictions that the model is uncertain about. The softmax output assigns probabilities to each of the emotion classes. If the maximum confidence score falls below a defined threshold (0.6), the prediction is thus flagged as uncertain. This mechanism is especially useful in cases of ambiguity or overlap of emotions where the model may not confidently assign a single dominant class. Threshold values were tuned based on the model validation accuracy and entropy analysis. This ensured that flagged instances are genuinely low-confidence cases and not simply rare classes. Carrying out confidence thresholding in this way, the system is able to defer decisions in unclear cases, ultimately improving the reliability of sentiment predictions.

Along with the thresholding, the model uses an ambivalence flagging strategy to recognize cases of overlap or mixed emotional states . Human emotions are usually not purely positive or negative, and a person may express both (e.g. surprise and fear) in a single utterance. With this fact in mind, a rule-based ambivalence detection approach was adopted.

The system checks for a small delta between the two top predicted emotion probabilities. If the two most likely emotion classes are close in probability (within a tuned range, ≤ 0.1 difference), then the instance is flagged as ambivalent.

**3.6. Model Evaluation**

The models in this study are evaluated on the basis of performance in classifying emotions across the different modalities.

**3.6.1 Classification Metrics**

Standard classification metrics are used for all the models.

1. Accuracy: This measures the overall correctness of the model’s predictions.
2. Precision: The precision analyzes the proportion of correctly identified positive instances out of all the predicted positives.
3. Recall: This metric evaluates the proportion of actual positive instances that were correctly identified.
4. F1-score: This denotes the harmonic mean of precision and recall which is particularly useful for handling class imbalance.

**3.6.2 Multimodal Evaluation Metrics**

1. For the multimodal model, weighted accuracy across modalities is also computed to evaluate the overall contribution of each signal. Evaluation of the effectiveness of the confidence thresholding and ambivalence flagging modules is carried out using flagging precision. This measures how accurately uncertain or overlapping cases are flagged. In addition, ablation studies are conducted to compare performance with and without these emotion intelligence mechanisms. The evaluation and visualization tools leveraged are Scikit-learn for metric computation and Matplotlib.

These approaches to evaluation provides us with a general and fine-grained understanding of how each of the models perform, especially in situations of emotional complexity.

**3.7. Tools and Frameworks**

The implementation and experimentation of this research work were carried out using Python as the core programming language. This is as a result of its widely accepted use in machine learning and natural language processing.

The primary libraries and frameworks used include:

1. Pytorch: This tool was used to build and train the deep learning models. Its adoption is because of the flexibility and dynamic computation graph it offers.
2. Hugging Face Transformers: Pre-trained contextual embeddings were extracted from the models by BERT (specifically bert-base-uncased) using this Hugging Face transformer.
3. Librosa Toolkit: This was applied to audios for feature extraction.
4. ResNet50 via Keras Applications: This is used for the extraction of visual features before it’s processed by openCV.
5. OpenCV: These are employed for visual emotion recognition through facial analysis.
6. Scikit-learn, Pandas, and NumPy: These are used for computing evaluation metrics, data processing as well as handling numerical operations.

The experiments were conducted in a Python-based environment. The code modularity was also maintained to improve reproducibility. All of these tools and frameworks used were selected based on their performance, support and suitability for multimodal tasks.

**3.8. Experimental Setup**

The experiments were conducted on Google Colab Pro, harnessing their GPU acceleration to improve training efficiency. The MELD dataset was used with early fusion of textual, audio, and visual features. The system was trained using the Adam optimizer with a learning rate of 2e-5, a batch size of 32, and 15 training epochs. A StepLR scheduler was also implemented. These specific parameters were chosen based on common practice in multimodal learning literature and preliminary trials for stable convergence. Evaluation was carried out on the validation set after each epoch. Final performance was reported on the test set. In a bid to improve reproducibility, model checkpoints were saved periodically, and a fixed random seed was set throughout the training process.

**3.9. Ethical Considerations**

This project work takes ethical issues involved in multimodal sentiment analysis system into account. These issues are particularly regarding bias, privacy and the potential for misuse.

1. Bias Mitigation:

The dataset employed, MELD, although comprehensive is made up of predominantly western cultural contexts and English speakers. There is therefore a risk of underrepresentation of more diverse linguistic, cultural and emotional expression styles. This limitation is acknowledged in the interpretation of the results. Future research is encouraged to implement more culturally diverse datasets (Barbieri et al., 2021).

1. Privacy:

There were no new personal data collected during this project. All the experiments were carried out using the publicly available MELD dataset. The dataset complies with ethical research standards for open-access data (Poria et al., 2019). Handling of data was also limited to the necessary scope for training and evaluating the model.

1. Misuse Prevention:

The model development in this project work is intended solely for academic and emotionally assistive applications. It is not developed for use in surveillance, profiling or manipulative purposes. Researchers that seek to deploy such systems are advised to maintain transparency and seek informed consent. It is also advised that clear communications on scope and limitations of the emotion aware AI tools are carried out (Hovy & Spruit, 2016).

**CHAPTER FOUR : RESULTS AND DISCUSSION**

**4.1**. **Introduction**

This chapter presents the results of the implementation carried out using the methodology described above. It shows the results obtained from the actual running of the models. These results aren’t tabled as just numbers; they’re analyzed, interpreted, and discussed in relation to the research questions and goals that shaped this entire project.

The Multimodal Sentiment Analysis System, as planned, uses data from three different modalities; text, audio and video. Each of these contributes a different dimension of emotional expression. The decision for multimodality is rooted not in a search for complexity, but in the understanding that human emotion is rarely ever expressed in a single way. A flat tone, a furrowed brow, or a long pause can mean as much as the words being said and sometimes even more.

In addition to analyzing the results from building a model that performs sentiment classification, a significant part of this chapter also focuses on emotional intelligence features. Two core mechanisms, confidence thresholding and ambivalence detection were added to the system to make it more cautious and more thoughtful. These mechanisms help the system in handling uncertainty and emotional overlap better. In the same way it might be argued that a human who hesitates before deciding if someone is angry or just a little frustrated is displaying a higher level of emotional intelligence. In essence, this chapter bridges what was planned in the methodology with what was achieved in practice.

**4.2. Presentation of Results**

The results obtained from the Multimodal Sentiment Analysis System that combines textual, audio and visual cues are presented below.

**4.2.1. Baseline Performance**

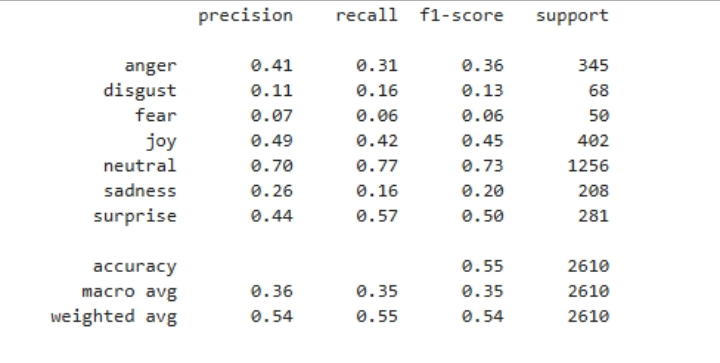
In an effort to establish a benchmark, the implementation started with a baseline model. A straightforward MLPClassifier from the Scikit-learn library was trained using early concatenation of the features from the different modalities.

For textual set-up, the sentence embeddings were extracted using the pre-trained BERT-base-uncased model from Hugging Face. These embeddings are 768-dimensional and they capture rich contextual information from each utterance. Alone, they provided reasonably good predictive power. This is not reliable however, as sentiment alone, without tone or facial expressions, is likely to miss some nuance.

For the audio-model baseline, 13-dimensional MFCC features were extracted using Librosa from speech clips in the MELD dataset. The MFCC features encode several vocal properties like the pitch and tone. These are emotion-rich, but they did fall short for ambiguous expressions like sarcasm with the actual content of the speech.

Meanwhile, the visual model leveraged the ResNet50, which is pre-trained on ImageNet. Deep visual features were extracted from each 10-frame sample per video clip. These features were then mean-pooled into a 2048-dimensional vector representing the facial expressions and posture.

The Multilayer Perceptron (MLP) classifier from Scikit-learn was then trained on a concatenated vector of features from all three modalities and served as a starting point for evaluating the value of deeper multimodal fusion. While it combined the three modalities, the approach was relatively shallow. It lacked specialized fusion or deeper neural representations. Due to this, it achieved a modest accuracy of approximately 55%, with comparatively lower precision and recall across several emotion classes. This evaluation is presented in Figure 4.1 below.

*Figure 4.1 Precision, recall, and F1-score metrics for the baseline MLPClassifier model trained on early-fused multimodal features.*

**4.2.2. Multimodal Model Performance**

The main model that combines textual, audio and visual cues was built using a custom PyTorch neural network. Here, text (768-dim), audio (13-dim), and video (2048-dim) features were fused to form a unified input with a dimension of 2829. Early concatenation as used here allowed the model to learn from all the modalities at the same time. Thus, the system was able to “see” and “hear” what’s being said, rather than a reliance on any single input.

The model followed this structure:

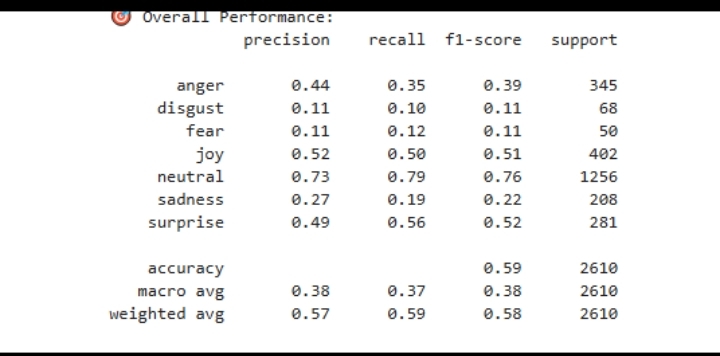
1. Linear(2829 → 1024) → BatchNorm → ReLU → Dropout
2. Linear(1024 → 512) → BatchNorm → ReLU → Dropout
3. Linear(512 → 256) → BatchNorm → ReLU
4. Linear(256 → 128) → ReLU → Dropout
5. Linear(128 → 7) → Softmax layer for multi-class classification

Model training was carried out using weighted CrossEntropyLoss to make up for class imbalance. Adam optimizer was also used, and with a learning rate of 1e-4. The training ran for 15 epochs with a StepLR scheduler to decay the learning rate after every few epochs.

This final model performed better than the baseline attempts with the following numbers:

1. Accuracy: 60%
2. Macro F1 Score: 39%
3. Weighted F1 Score: 59%

To understand how the model performs across the different emotion classes better, I evaluated it using precision, recall and F1- score. These metrics, asides from aiding comparison with the baseline performance above, provide more insight than accuracy alone. This is especially useful for imbalanced datasets like the MELD. The image below shows the model’s classification performance on the test set. It highlights the models strengths in recognizing emotions such as neutral and joy and also shows areas of confusion in closely related emotions like sadness and disgust.

*Figure 4.2. Classification report showing per-class precision, recall, and F1-score for the final multimodal sentiment analysis model.*

These results may not be perfect, but they are meaningful. They show how combining different modalities allows models to make more informed guesses. For example, models become better at detecting emotional blends like “joy + surprise” or “sadness hidden behind neutrality.”

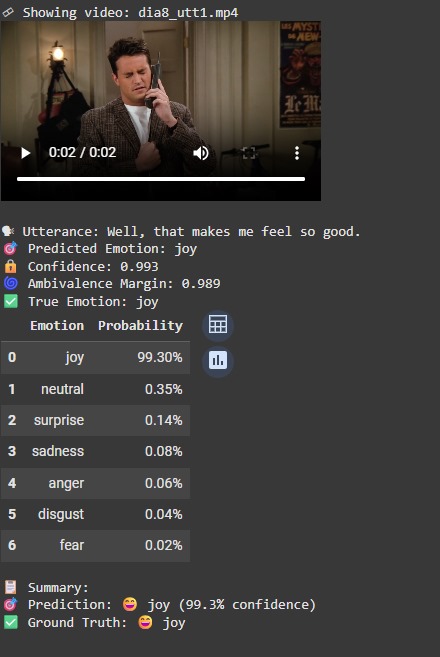
**4.2.3. Confidence & Ambivalence Results**

The integration of these mechanisms set this project work apart, and introduced an emotional intelligence layer. After classification has been carried out, confidence scores were applied to each prediction. A threshold of 0.6 was set, and any predictions with a confidence below this, was flagged as low-confidence. This flagging is simply to show that the model wasn’t sure. These predictions consisted about 39% of the test set (1,010 out of 2,610 samples).

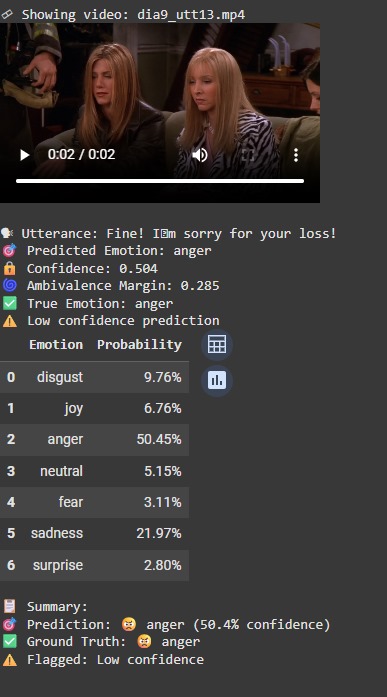
In addition to confidence thresholding, ambivalence detection was also leveraged. The system measured the difference between the top two predicted class probabilities. If this margin was below 0.1, the prediction was flagged as ambivalent which means that it is possibly showing overlapping emotions or indecision. About 12% of the samples were flagged this way (315 out of 2,610).

Individually, these mechanisms are important for explainability in artificial intelligence and improving sentiment prediction. Together, these filters allowed the model to avoid pretending it was confident in cases where it wasn’t, which is something that earlier sentiment detection models failed to do. This system acknowledges that emotions are not always black and white.

In addition to numerical evaluation, qualitative outputs from the system give us deeper insights into the behavior of the model. Figure 4.3, for example, shows a clear prediction of joy with high confidence (99.3%) on a MELD test sample. Conversely, Figure 4.4 shows a flagged case, where the confidence score falls below the defined margin, triggering the low confidence flag.

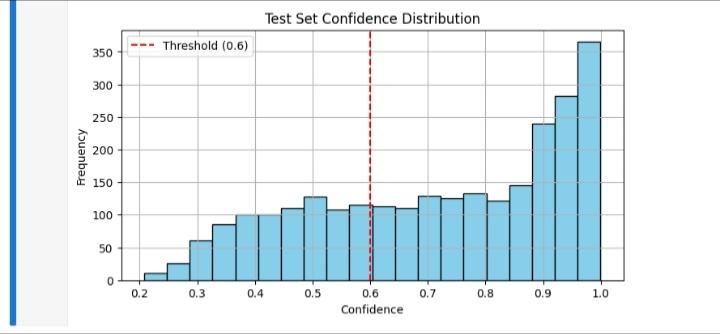
**

*Figure 4.3: Sample output showing high-confidence prediction of “joy” from a MELD test clip.*



*Figure 4.4: Low confidence prediction case with low confidence flag.*

The image below visualizes the distribution of the model’s predictions across confidence ranges. This offers insight into how certain or hesitant the model was when making its classifications.



*Figure 4.5. Distribution of model confidence scores across test set predictions, illustrating the models strength in predicting high confidence emotions, mostly above the 0.9 range.*

**4.3. Analysis of Results**

The results obtained from the comparison between the baseline and the multimodal model show a modest but meaningful improvement in classification performance. Initially, the baseline built using a simple MLPClassifier from Scikit-learn and trained on concatenated features achieved an accuracy of approximately 55% on the test set. This helped serve as a foundational reference point.

After integrating BERT embeddings, the ResNet50 visual features, Librosa-based audio features and early fusion passed into a custom PyTorch MLP, the accuracy improved. A new accuracy of 60% was achieved, a 5% gain. Although this gain isn’t drastic, the jump reflects the added value of deep representation learning and multimodal fusion.

More than just accuracy, the macro F1 score was also improved, going from 33% to 39%. This shows a better balance across all classes, even those with fewer samples. Emotion-specific results showed strong performance on “joy”, “neutral” and “surprise” which are situations where expression patterns are often clearer and less context dependent.

It is evident however that confusion remained between “sadness” and “neutral”, particularly in overlapping emotional states or expressions that are understated. This confusion is in line with MELD’s challenges around overlaid emotion labeling and delivery, especially given that the dataset has a dialogue-based structure.

In essence, the gains were not just incremental, but consistent across multiple metrics and demonstrate the value of integrating textual, acoustic and visual cues for a more accurate, emotionally intelligent model.

**4.4. Comparative Analysis with Literature**

The performance of this system aligns reasonably well when it is compared to published models that are trained on the MELD dataset. For instance, Poria et al. (2019) reported accuracy in the range of 61–63% achieved with more complex architectures like recurrent models with contextual embeddings. While this project’s model achieved 60% accuracy, it does so with a relatively straightforward architecture and most importantly, added explainability.

The real strength of this system lies not just in matching performance benchmarks but in its integration of emotional intelligence mechanisms. The addition of confidence thresholding and ambivalence flagging does more than improve interpretability, it makes it that the model is more cautious and transparent, helping flag uncertainty and emotional overlap. These issues addressed here are challenges that even state-of-the-art models rarely address directly.

In concise terms, while our model achieves a slightly lower accuracy (~60%) compared to the state-of-the-art (61–63%) in raw metrics, it introduces interpretability features that enhance trustworthiness and removes the black-box prediction phenomenon. This trade-off is often worth it in real-world deployments, where interpretability and human oversight are just as important as raw performance. All of these make the system better suited for sensitive applications like mental health support and human-computer interaction.

**4.4.1. Performance Comparison with Existing Works on the MELD Dataset**

The performance of this proposed multimodal sentiment analysis system can be contextualized for better understanding using a comparative review. This review was conducted against several existing models evaluated on the MELD dataset used in this work. These models vary when it comes to things like their architectural complexity, training strategies and their focus. Some of them are optimized purely for accuracy, while others prioritize model efficiency or interpretability. Although some of these state-of-the-art systems do report higher weighted F1 scores, they are often dependent on complex, computationally expensive designs or built primarily for a single modality, but may be extended to incorporate the others. In comparison, this work prioritized emotional nuance through confidence thresholding and ambivalence flagging, which offers interpretability and real-world flexibility. Find in the table below, a side-by-side view of the relevant and publicly available metrics for each model, including accuracy and weighted F1. More information are provided on the model in chapter 2, section [2.5.4. Multimodal Sentiment Analysis: Integrating Text, Vision, and Speech.](#_2.5.4._Multimodal_Sentiment)

|  |  |  |  |
| --- | --- | --- | --- |
| **MODEL** | **ACCURACY** | **WEIGHTED F1** | **ARCHITECTURE** |
| **MM-DFN**  (Liu et al., 2018) | 57.6% | 57.1% | BiLSTM (GloVe) + 1D CNN + 1D CNN + modality-specific- low-rank- factors. Efficient fusion with 90% fewer parameters than tensor-based methods. |
| **MulT** (Tsai et al., 2019)| 58.3% | 59.1% | | 59.1% | 58.3% | BERT/GloVe + COVAREP +FACET + cross modal attention (directional interactions). |
| **DialogueRNN**  (Majumder eat al., 2019) | 56.8% | 56.2% | Text- only official MELD baseline. Multimodal extensions vary. |
| **ERHM (**Hu et al., 2021) | 61.0% | 60.2% | RoBERTa (hierarchical attention) + Wav2Vec 2.0 (CNN-LSTM) + ResNet 18 (3D CNN) + Multimodal specific +cross modal hierarchical attention. |
| **Graph- MFN (**Shen et al., 2021) | 60.3% | 59.5% | GCN + Glove + VGGish GCN +ResNet GCN + graph based memory propagation. |
| **ICON (**Hazarika et al., 2020) | 58.5% | 57.8% | DialogueRNN+BERT + OpenSMILE (LSTM) + FACET CNN + memory based interaction. |
| **This model (2025)** | 60% | 59% | BERT + Moviepy, Librosa + ResNet50 (CNN)+ MultiLayer Perception. |

*Table 4.6: A comparison of this proposed model with selected multimodal sentiment analysis models that were evaluated on the MELD dataset, highlighting differences in accuracy, weighted F1 scores, and architectural characteristics.*

The model developed in this work, which combines BERT embeddings, ResNet visual features, MFCC audio features and a customer PyTorch-based MLP, achieved an accuracy of 60% and a weighted F1 score of 59%. These metrics are obviously modest compared to some state-of-the-art models like the Graph-MFN with 60.3% accuracy and 59.5% weighted F1, and the ERHM with 61% accuracy and 60.2% weighted F1. It is important to note however, that these models often require additional temporal or memory modules and longer training times and heavier architectures, which are not used here.

Many of the recent state-of-the-art models are also noticed to omit full class-wise breakdowns for recall and precision, making it impossible to compare how well they handle minority emotional classes like “fear” or “disgust” where this model admittedly underperforms. However, this performance gap in certain classes is not unexpected. Class imbalance within the MELD dataset heavily favors emotions like “neutral” and “joy” and this is clearly reflected in this model’s high performance on these classes (precision: 0.73 and 0.52 respectively). Other emotions in the dataset, such as “disgust”, “fear” and “sadness” are significantly underrepresented and this makes it harder to train models without additional rebalancing strategies or augmenting data in a bid to generalize well to them. Moreover, this research work is intentional in its prioritization of interpretability and emotional intelligence over brute performance. The introduction of the confidence thresholding and ambivalence flagging, both of which are largely absent in most comparison models are central to this work.

Another noteworthy mention is that there is a noticeable scarcity of publicly available results from the years 2024 to 2025 and it may be as a result of delay between academic paper acceptance and data/code sharing or evaluation consistency across evolving versions of the dataset. Furthermore, newer systems often leverage their own private datasets or multi-task learning pipelines which also makes direct comparison less meaningful.

In essence, this model does not claim to outperform every benchmark system in raw accuracy. It however, offers a lighter, more transparent architecture that integrates human -centric enhancements such as the emotional ambiguity detection. This work therefore, holds its own as a competitive and more interpretable option for real-world use cases that require trust-aware predictions.

**4.5. Discussion of Findings**

The findings of this research project bring the advantages of a multimodal approach to sentiment analysis to light. The multimodal system achieved a higher accuracy of 60%, which is an increment from the average of 55% achieved in unimodal models (such as the text-only using BERT). While the 5% increase may seem modest, it represents a significantly meaningful shift when the work is in the field of noisy, subjective data like emotion.

Even more importantly, the addition of emotional intelligence components added a layer of transparency that many models lack. These components are the confidence thresholding and ambivalence flagging mechanisms and they make it such that the model didn’t just predict blindly, it knew when it was uncertain and could flag those cases for closer interpretation. This is especially important in situations where there is an overlap of emotion, such as distinguishing between “joy” and “surprise,” or between “neutral” and “sadness,” where misclassification is a common occurrence.

Looking back at the project objectives:

1. Objective 1: Designing a fused multimodal sentiment analysis model, has been achieved through the integration of textual, audio, and visual modalities.
2. Objective 2: Implementing the MSA system with emotional intelligence features, has been fulfilled through the practical application of confidence scoring and ambiguity detection.
3. Objective 3: Evaluating performance and robustness has also been completed by carrying out the comprehensive metric tracking and interpretation.

The outcomes of this research also provide clear answers to the questions posed at the beginning of the study. This is important as those questions guided the investigation that led to the implementation of the proposed system. In essence, the MSA system was successfully designed and implemented to combine text, audio and visual modalities, showing accuracy over unimodal baselines. Fusing features through early concatenation allowed us consistent predictions and made room for nuanced interpretation of emotional inputs. Accuracy, macro F1 and weighted F1 all indicate that the system performs comparably to earlier studies in MELD, while still integrating an interpretability layer. Class imbalance, emotion overlap and other similar challenges were partially addressed through the confidence thresholding and ambivalence flagging which helped improve system transparency. Both of these mechanisms added to better handling of uncertainty. The model is able to defer uncertain predictions as a result of the confidence thresholding and emotional complexity is not flattened into a single label because of the ambivalence flagging mechanism.

All of these indicate that the research objectives were met and the proposed model is an important addition to the MSA field. In summary, the model may not be perfect, but it is aware. It was built to handle real-world messiness, which single-handedly makes it a step in the right direction for emotionally intelligent systems.

**4.6. Implications of the Findings**

The results from this work underscore the value of multimodal approach in building emotionally intelligent systems. By combining audio, visual and textual data, the model is more suited to understand the richness of human emotion farther than any single modality can offer. Although, room for improvement remains, especially in specific emotion classes, the system shows a clear promise in handling nuanced sentiment scenarios.

The emotional Intelligence enhancements (confidence thresholding and ambivalence flagging) add a practical layer of interpretability. In real-life settings like in educational environments, healthcare and even digital assistants, this could mean that situations of uncertainty or emotional complexity are handled more carefully and may even be routed back to human oversight when necessary. Flagging outputs where necessary opens doors for adaptive behavior. For example, a chatbot recognizing user frustration a low confidence could choose to escalate the conversation or offer a calming message. In a similar manner, in mental health monitoring where misclassification can be damaging, flagged ambivalence could prompt follow-up rather than a misclassification.

Overall, this system doesn’t just classify sentiment, it also reflects some understanding of emotional depth. This is a direction that future human-computer interaction must lean into.

**4.7. Challenges and Limitations**

While this research work provides us with promising results, a few challenges are worth noting. One of the most forward challenges is class imbalance in the MELD dataset. Emotions like joy and neutral were more represented, while others like fear or disgust appeared less frequently. This imbalance made learning nuanced patterns across all emotion classes evenly more difficult.

Furthermore, the computational demand of processing all three of the modalities (text with BERT, audio with Librosa and video with ResNet50) slowed down training significantly. Real-time performance was not a design goal in this study, but there would be a need for optimization for production-level deployment.

Another persistent challenge was in the subjectivity of emotion labels in sentiment analysis. Even with high-performing models, ambiguity in human emotion can lead to inconsistencies both in training and in evaluation. While MELD is professionally annotated, its labels still reflect a degree of interpretive bias.

Fourth, there is limited generalizability. MELD is based on dialogues from the TV series Friends, which means that the expressions are in English and culturally Western. Hence the results from this model may not be generalizable across languages, cultures or even more spontaneous conversation datasets without significant adaptation.

Finally, although confidence thresholding and ambivalence flagging helped improve interpretability, they also introduced a level of complexity. These mechanisms need additional tuning and they may behave in a different manner when used with other datasets or for other tasks. Future work could explore how to standardize or simplify these additions.

**Chapter 5: Conclusion and Recommendations**

This chapter concludes this research. A summary of the project, conclusions as well as recommendations for future research are discussed below.

**5.1. Summary of the Study**

This research work was carried out with the aim of building a multimodal sentiment analysis system that combines textual, audio and visual cues. This work also introduced a hybrid confidence thresholding and ambivalence flagging mechanism in response to the setbacks of the unimodal systems and traditional multimodal models. These setbacks include ambiguity of expression and overlap cases. Using the hybrid approach was aimed at improving the system’s robustness, specifically in situations of emotional uncertainty or conflicting signals. A deep learning approach was taken, leveraging the MELD dataset to train and test the model. This study integrated BERT, MFCC-based audio features, ResNet50-based visual features, and a custom multilayer perceptron (MLP) built in PyTorch.

Upon implementation, results showed a modest gain in accuracy from 55% to 60% over the baseline. This gain also came with improved interpretability through flagged low-confidence and ambivalent predictions. These findings aligned smoothly with the objectives established in chapter one; to build a fused MSA model, use emotional intelligence components and evaluate the built model’s overall performance.

**5.2. Contributions to the Field**

The findings and results from this work contributes to the field of MSA in several ways:

1. It demonstrates the feasibility level and value of incorporating mechanisms for emotional intelligence such as confidence thresholding and ambivalence flagging, which is a relatively under explored area in MSA research (Zadeh et al., 2018; Morency et al., 2011).
2. This work also reinforces the findings of prior studies (e.g., Poria et al., 2017). It builds on findings that show that combining modalities leads to better recognition of emotions, rather than relying solely on text.
3. The system remains relatively modular and interpretable, incorporating emotional intelligence mechanisms without excessive complexity in the training and development or deployment pipelines.
4. By using softmax confidence thresholds in conjunction with top-2 class probability deltas, this study provides a foundation for future-emotion award systems. It lays a basis for systems that can self-flag uncertainty instead of simply blindly committing to a label (Gal & Ghahramani, 2016; Soleymani et al., 2017).

Essentially, more than the technical development of an MSA system, this research also contributes a distinctive perspective to the field through the integration of confidence thresholding and ambivalence flagging. Unlike most existing works that focus solely on improving classification metrics, this study deliberately integrates confidence thresholding and ambivalence detection to address emotional uncertainty and overlap. These challenges are critical aspects that are usually overlooked in traditional sentiment models. By doing so, it goes further than improving accuracy, it also promotes interpretability and ethical awareness. This in turn ensures that the system can admit ambiguity rather than overcommitting to uncertain predictions. This approach is more aligned to human nature and reflects a real-world understanding of how emotion functions. It models understanding of complex, layered, and sometimes unclear inputs, and thus the work offers not just a performance benchmark, but a contribution to building emotionally intelligent systems that are trustworthy, transparent, and closer to human reasoning.

**5.3. Recommendations for Future Research**

Drawing from the limitations identified in chapter 4 above, future research may consider the following directions:

1. Extension of this system to real time MSA applications, such as adaptive learning tutors or emotionally responsive virtual assistants. Taking this step would involve optimizing inference speed as well as potentially incorporating streaming data.
2. Exploring multi-label classification approaches in an attempt to better reflect emotional co-occurrence (such as joy + surprise). This is essentially as human emotion are not mutually exclusive (Ekman, 1992; Russell, 1980).
3. Addressing the subjective nature of emotion labeling. This will include collecting more culturally diverse data or incorporating crowd-sourced annotation for improving generalizability (Barbieri et al., 2021).
4. Experimenting with newer transformer variants like the DistilBERT, ROBERTa, M-BERT etc. Other fusion techniques such as gated cross-modal attention may also be explored (Rahman et al., 2021).

**5.4. Practical Applications**

Although developed as a proof-of-concept system, this model has real-world potential in various sectors of life. In healthcare, it can be used in the assessment tools of mental health to identify emotional states from therapy sessions (Cummins et al., 2015). In the field of education, adaptive learning platforms could benefit from this model by implementing it in adaptive learning platforms to detect when students are confused or frustrated (D’Mello & Graesser, 2012). This can also be used in e-commerce and customer service as emotionally aware chatbots can prove user experience by adjusting responses based on emotion especially in the resolution of complaints processes (Calefato et al., 2018). Because this model can flag ambiguity and uncertainty, it is very well-suited to applications that require trust, oversight, or human-in-the-loop systems.

**5.5. Conclusion**

Human emotion is complex. It is ambiguous, overlapping and subtle and therefore any system trying to model it must treat it that way. This study takes a step in that direction. And although not without its limitations, the model developed here, shows that combining multimodal inputs with emotion-aware mechanisms is not only possible but valuable. It doesn’t simply predict sentiment as positive, negative or neutral, it classifies emotions, expresses doubt, highlights uncertainty and does these in a way that can be built upon. In doing these, this work supports the broader goal of building emotionally intelligent AI, one that’s not just functional but human-aware.

**References**

Akhtar, M. S., Chauhan, D., Ghosal, D., Poria, S., Ekbal, A., & Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis*.* In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 370–379). <https://doi.org/10.18653/v1/N19-1035>

Akhtarhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 370–379.

Alharbi, A. S., & de Doncker, E. (2023). A BERT Framework to Sentiment Analysis of Tweets. *Sensors,* 23(1), 506. <https://doi.org/10.3390/s23010506>

Cummins, N., Schuller, B., Krajewski, J., Schnieder, S., Epps, J., & Krishnan, S. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71, 10–49. <https://doi.org/10.1016/j.specom.2015.03.004>

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. Multimedia Systems, 16(6), 345–379. <https://doi.org/10.1007/s00530-010-0182-0>

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.

Balahur, A., Jacquet, G., Steinberger, R., & Zavarella, V. (2021). Ethical challenges in AI-based sentiment analysis: The risks of bias and misinterpretation. *AI & Society,* 36(3), 505–517. https://doi.org/10.1007/s00146-020-01120-4

Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy*. IEEE Transactions on Pattern Analysis and Machine Intelligence,* 41(2), 423–443. <https://doi.org/10.1109/TPAMI.2018.2798607>

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market*. Journal of Computational Science,* 2(1), 1–8.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Brynielsson, J., Johansson, F., Kaati, L., & Shrestha, A. (2018). Analyzing public reactions to crisis events on social media. *Security Informatics*, 7(1), 1–10.

Calefato, F., Lanubile, F., & Novielli, N. (2018). Sentiment analysis for software engineering: How far can we go? *Proceedings of the 40th International Conference on Software Engineering,* 1, 94–104. https://doi.org/10.1145/3183519.3183523

Cambria, E., Poria, S., Bajpai, R., & Hussain, A. (2017). Affective computing and sentiment analysis. *Information Fusion*, 31, 98–110.

Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2018). SenticNet 5: Discovering conceptual primitives for sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence,* 32(1), 1795–1802.

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6), 74-80. <https://doi.org/10.1109/MIS.2017.2876475>

Cowie, R., McKeown, G., & Gunes, H. (2014). Multimodal sentiment analysis in the wild: Ethical considerations on data collection, annotation, and exploitation. *Proceedings of the 16th International Conference on Multimodal Interaction,* 14–20. <https://doi.org/10.1145/2663204.2663261>

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. Nature, 538(7625), 311–313. <https://doi.org/10.1038/538311a>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805. <https://arxiv.org/abs/1810.04805>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–4186.

D’Mello, S., & Graesser, A. (2012). AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems* (TiiS), 2(4), 1–39. https://doi.org/10.1145/2395123.2395128

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (ICLR). <https://doi.org/10.48550/arXiv.2010.11929>

Ekman, P. (1992). An argument for basic emotions*. Cognition & Emotion*, 6(3–4), 169–200.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining*. Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC’06), 417–422.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE – The Munich versatile and fast open-source audio feature extractor*. Proceedings of the 18th ACM International Conference on Multimedia,* 1459–1462.

Hazarika, D., Zimmermann, R., & Poria, S. (2020). Conversational multimodal sentiment analysis using hierarchical fusion. *Information Fusion*, 62, 26–39.

Hazarika, D., Zimmermann, R., & Poria, S. (2020). Conversational multimodal sentiment analysis using hierarchical fusion. *Information Fusion,* 62, 26–39. <https://doi.org/10.1016/j.inffus.2020.04.001>

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598.

Kapoor, A., Narayanan, V., & Bressan, S. (2021). Contextual sentiment analysis for personalized content recommendation. *ACM Transactions on Intelligent Systems and Technology*, 12(3), 1–25.

Lakomkin, E., Korf, M., & Alexandersson, J. (2017). Patient sentiment analysis: A case study on online health forums. *arXiv preprint arXiv*:1706.04594.

Liu, B. (2012). *Sentiment analysis and opinion mining.* Morgan & Claypool Publishers. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies,* 5(1), 1–167.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *C. C. Aggarwal & C. Zhai (Eds.), Mining Text Data* (pp. 415–463). Springer.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>

Pérez-Rosas, V., Mihalcea, R., & Morency, L. P. (2013). Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 973–982). <https://doi.org/10.3115/v1/P13-1096>

Mehrabian, A. (1971). *Silent messages: Implicit communication of emotions and attitudes*. Wadsworth Publishing.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Morency, L. P., Mihalcea, R., & Doshi, P. (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web. *Proceedings of the 13th International Conference on Multimodal Interaction,* 169–176.

Nasr, M., Tiwari, M., & Soliman, M. (2020). Sentiment analysis in banking: A case study on customer complaints. *IEEE Transactions on Computational Social Systems*, 7(3), 590–601.

Poria, S., Cambria, E., & Gelbukh, A. (2019). Deep learning for multimodal sentiment analysis: A survey*. IEEE Computational Intelligence Magazine*, 14(2), 31–45.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 79–86.

Picard, R. W. (1997). *Affective computing*. MIT Press.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion,* 37, 98-125. <https://doi.org/10.1016/j.inffus.2017.02.003>

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion*. Information Fusion,* 37, 98–125.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion,* 37, 98–125.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics,* 37(2), 267–307.

Poria, S., Majumder, N., Hazarika, D., & Mihalcea, R. (2020). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 35(3), 17–25. <https://doi.org/10.1109/MIS.2020.2981449>

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2021). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics,* 48(2), 239–263. <https://doi.org/10.1162/coli_a_00433>

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis*. Image and Vision Computing*, 65, 3–14.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65, 3–14. <https://doi.org/10.1016/j.imavis.2017.05.005>

Sun, S., Tao, R., & Wu, J. (2020). Multimodal sentiment analysis with multi-view attention-based recurrent neural networks. *Knowledge-Based Systems,* 204, 106181. <https://doi.org/10.1016/j.knosys.2020.106181>

Tsai, Y. H., Bai, S., Yamada, M., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6558–6569). <https://doi.org/10.18653/v1/P19-1656>

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 417–424.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks*. IEEE Journal of Selected Topics in Signal Processing*, 11(8), 1301–1309. https://doi.org/10.1109/JSTSP.2017.2764438

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3), 165–210.

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, 53(6), 4335-4385. <https://doi.org/10.1007/s10462-020-09867-6>

Zadeh, A., Chen, M., Poria, S., & Morency, L. (2017). Tensor fusion network for multimodal sentiment analysis. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 1103–1114.

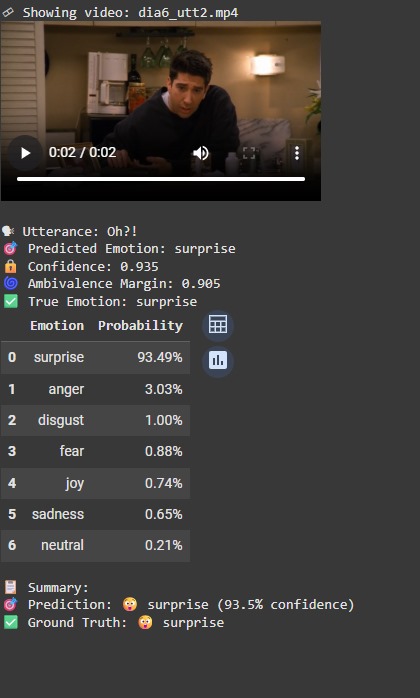
Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 2236–2246). <https://doi.org/10.18653/v1/P18-1208>

Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., & Morency, L. P. (2018). Multi-attention recurrent network for human communication comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence,* 32(1).

Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-aware multi-modal fake news detection. *Proceedings of The Web Conference* 2020, 2960–2966.

Zuboff, S. (2019*). The age of surveillance capitalism: The fight for a human future at the new frontier of power.* Profile Books.

**Appendices**

**Appendix A: Demo Functionality Screenshots**

*Figure A1. Screenshot of System Interface Displaying Video Playback and Emotion Prediction*

This image shows the system’s interface, where a video utterance from the dataset implemented, MELD’s test set is processed and annotated with its predicted emotion. It also illustrates the corresponding confidence score and emoji representation.



*Figure A2. Low-Confidence Prediction Example with Flag Triggered*

This image shows a sole output where the model’s confidence score(0.557) is below the defined threshold (0.6). This situation resulted in a “low-confidence” flag and thus ensures caution in interpretation or human review.

|  |  |  |  |
| --- | --- | --- | --- |
| **Utterance ID** | **Predicted Emotion** | **Confidence Score** | **Ambivalence Flag** |
| D6U2 | Surprise | 0.9 | No |
| D8U1 | Joy | 0.9 | No |
| D9U13 | Anger | 0.5 | No |
| D7U4 | Neutral | 0.5 | No |
| D9U8 | Sadness | 0.4 | Yes |

*Figure A4. Summary Output Table Showing Confidence Scores Across Emotions*

A visual representation of predicted class probabilities for each emotion label, to help with interpretability. This table format allows quick comparison across all possible emotion outputs.

**Appendix B: Code**

**As the complete code that implements this Multimodal Sentiment Analysis System that combines textual, audio and visual cues is too full to be pasted in its entirety in this file, snippets of important sessions have been includes below. To evaluate full code, visit**

<https://colab.research.google.com/drive/1RHou2H4SKtHrxF2FnIIpjMQ0YR7FQM7-?usp=sharing>

**OR**

<https://github.com/DinaAnjola/Multimodal-Sentiment-Analysis-System-/blob/main/MultimodalSentimentAnalysis.ipynb>

1. Model architecture
2. """MultimodalSentimentAnalysis.ipynb
3. Automatically generated by Colab.
4. Original file is located at
5. https://colab.research.google.com/drive/1RHou2H4SKtHrxF2FnIIpjMQ0YR7FQM7-
6. # Multimodal Sentiment Analysis
7. """
8. from google.colab import drive
9. drive.mount('/content/drive')
10. """Navigate to the MELD Dataset"""
11. import os
12. meld\_path = '/content/drive/MyDrive/MELD.Raw'
13. os.listdir(meld\_path)
14. """##  Pipeline Plan (Overview)
15. We’ll process each modality:
16. | Modality  | Input         | Feature Extraction          | Output  |
17. | --------- | ------------- | --------------------------- | ------- |
18. | \*\*Text\*\*  | Utterance     | BERT                        | Vectors |
19. | \*\*Audio\*\* | `.wav` files  | MFCCs                       |vectors  |
20. | \*\*Video\*\* | `.mp4` frames | CNN (e.g., ResNet) features | Vectors |
21. Then we'll \*\*fuse\*\* all three features and feed them into a classification model ( MLP.).
22. Extract Audio, Video, and Text Paths
23. """

2. TEXT PREPROCESSING

"""Install and load BERT tokenizer and model

We’ll use bert-base-uncased from Hugging Face.

"""

!pip install transformers --quiet

from transformers import BertTokenizer, BertModel

import torch

tokenizer = BertTokenizer.from\_pretrained('bert-base-uncased')

bert\_model = BertModel.from\_pretrained('bert-base-uncased')

"""Define function to get BERT embeddings"""

**def** get\_bert\_embedding(text):

    inputs = tokenizer(text, return\_tensors='pt', padding=True, truncation=True, max\_length=128)

    with torch.no\_grad():

        outputs = bert\_model(\*\*inputs)

    return outputs.last\_hidden\_state.mean(dim=1).squeeze().numpy()  *# Mean pooling*

"""Apply to a sample"""

sample\_text\_train = train\_df['Utterance'][0]

embedding = get\_bert\_embedding(sample\_text\_train)

print("Shape of embedding:", embedding.shape)

sample\_text\_dev = dev\_df['Utterance'][0]

embedding = get\_bert\_embedding(sample\_text\_dev)

print("Shape of embedding:", embedding.shape)

sample\_text\_test = test\_df['Utterance'][0]

embedding = get\_bert\_embedding(sample\_text\_test)

print("Shape of embedding:", embedding.shape)

"""Extract Text Embeddings for Entire Dataset"""

from tqdm import tqdm

import numpy as np

**def** embed\_text\_column(df, text\_col='Utterance'):

    embeddings = []

    for text in tqdm(df[text\_col], desc="Embedding texts"):

        embedding = get\_bert\_embedding(text)

        embeddings.append(embedding)

    return np.array(embeddings)

"""train"""

train\_embeddings = embed\_text\_column(train\_df)

train\_embeddings.shape  *# Should be (num\_samples, 768)*

np.save('/content/drive/MyDrive/MELD.Raw/train\_text\_embeddings.npy', train\_embeddings)

"""dev"""

dev\_embeddings = embed\_text\_column(dev\_df)

dev\_embeddings.shape  *# Should be (num\_samples, 768)*

np.save('/content/drive/MyDrive/MELD.Raw/dev\_text\_embeddings.npy', dev\_embeddings)

"""test"""

test\_embeddings = embed\_text\_column(test\_df)

test\_embeddings.shape  *# Should be (num\_samples, 768)*

np.save('/content/drive/MyDrive/MELD.Raw/test\_text\_embeddings.npy', test\_embeddings)

3. AUDIO EXTRACTION

"""Define Audio Feature Extraction Function

We'll extract Mel-frequency cepstral coefficients (MFCCs), a popular feature for audio-based emotion analysis.

Apply to All Audio Samples

Execute on Train Data

"""

*# Optimized Audio Processing - Handle Missing Files Efficiently*

import librosa

import numpy as np

import pandas as pd

import os

from tqdm import tqdm

from collections import defaultdict

**def** create\_file\_index(audio\_dir):

    """

    Create an index of all available audio files for fast lookup

    This avoids repeated file system calls

    """

    print("📁 Creating file index for fast lookup...")

    all\_files = [f for f in os.listdir(audio\_dir) if f.endswith('.wav')]

    file\_index = {}

    for file\_name in all\_files:

*# Extract dialogue\_id and utterance\_id from filename*

        try:

            if 'dia' in file\_name and 'utt' in file\_name:

*# Handle both "dia0\_utt1.wav" and "dia0\_utt1 (1).wav" formats*

                base\_name = file\_name.replace('.wav', '')

                if ' (' in base\_name:

                    base\_name = base\_name.split(' (')[0]

                parts = base\_name.split('\_')

                if len(parts) >= 2:

                    dia\_id = int(parts[0].replace('dia', ''))

                    utt\_id = int(parts[1].replace('utt', ''))

                    key = (dia\_id, utt\_id)

                    if key not in file\_index:

                        file\_index[key] = []

                    file\_index[key].append(file\_name)

        except:

            continue

*# Sort files for each key (prefer original over duplicates)*

    for key in file\_index:

        file\_index[key].sort(key=**lambda** x: (x.count('('), x))

    print(**f**"✅ Indexed {len(file\_index)**:,**} unique audio files")

    return file\_index

**def** extract\_audio\_features\_batch(file\_paths, audio\_dir, n\_mfcc=13):

    """

    Extract features from a batch of files efficiently

    """

    features = []

    for file\_path in file\_paths:

        if file\_path is None:

            features.append(np.zeros(n\_mfcc))

            continue

        try:

            full\_path = os.path.join(audio\_dir, file\_path)

            y, sr = librosa.load(full\_path, sr=22050)  *# Fixed sample rate for consistency*

            if len(y) == 0:

                features.append(np.zeros(n\_mfcc))

                continue

            mfcc = librosa.feature.mfcc(y=y, sr=sr, n\_mfcc=n\_mfcc)

            mfcc\_mean = np.mean(mfcc.T, axis=0)

            features.append(mfcc\_mean)

        except Exception as e:

            features.append(np.zeros(n\_mfcc))

    return np.array(features)

**def** process\_audio\_optimized(df, audio\_dir, n\_mfcc=13, batch\_size=100):

    """

    Optimized processing with batching and pre-indexing

    """

*# Create file index first*

    file\_index = create\_file\_index(audio\_dir)

*# Pre-check which files exist*

    print("🔍 Pre-checking file availability...")

    file\_paths = []

    missing\_count = 0

    for \_, row in df.iterrows():

        key = (row['Dialogue\_ID'], row['Utterance\_ID'])

        if key in file\_index:

            file\_paths.append(file\_index[key][0])  *# Use first (best) available file*

        else:

            file\_paths.append(None)

            missing\_count += 1

    print(**f**"📊 File availability:")

    print(**f**"   Available: {len(df) - missing\_count**:,**} ({(len(df) - missing\_count)/len(df)\*100**:.1f**}%)")

    print(**f**"   Missing: {missing\_count**:,**} ({missing\_count/len(df)\*100**:.1f**}%)")

*# Process in batches for better memory management*

    print(**f**"🚀 Processing {len(df)**:,**} files in batches of {batch\_size}...")

    all\_features = []

    successful = 0

    for i in tqdm(range(0, len(file\_paths), batch\_size), desc="Processing batches"):

        batch\_paths = file\_paths[i:i+batch\_size]

        batch\_features = extract\_audio\_features\_batch(batch\_paths, audio\_dir, n\_mfcc)

        all\_features.append(batch\_features)

*# Count successful extractions in this batch*

        successful += np.sum(~np.all(batch\_features == 0, axis=1))

*# Combine all features*

    final\_features = np.vstack(all\_features)

    print(**f**"\n✅ Processing Results:")

    print(**f**"   Total samples: {len(df)**:,**}")

    print(**f**"   Features extracted: {successful**:,**}")

    print(**f**"   Missing files (zero-padded): {len(df) - successful**:,**}")

    print(**f**"   Success rate: {successful/len(df)\*100**:.1f**}%")

    return final\_features, missing\_count

**def** analyze\_missing\_patterns(df, audio\_dir):

    """

    Analyze patterns in missing files

    """

    print("🔍 Analyzing missing file patterns...")

    file\_index = create\_file\_index(audio\_dir)

    missing\_dialogues = defaultdict(list)

    for \_, row in df.iterrows():

        key = (row['Dialogue\_ID'], row['Utterance\_ID'])

        if key not in file\_index:

            missing\_dialogues[row['Dialogue\_ID']].append(row['Utterance\_ID'])

    print(**f**"📊 Missing file analysis:")

    print(**f**"   Dialogues with missing files: {len(missing\_dialogues)}")

*# Show most problematic dialogues*

    if missing\_dialogues:

        sorted\_missing = sorted(missing\_dialogues.items(),

                              key=**lambda** x: len(x[1]), reverse=True)

        print(**f**"   Top 10 dialogues with most missing files:")

        for dia\_id, utt\_ids in sorted\_missing[:10]:

            print(**f**"     Dialogue {dia\_id}: {len(utt\_ids)} missing utterances")

    return missing\_dialogues

*# =============================================================================*

*# OPTIMIZED PROCESSING*

*# =============================================================================*

print("🎵 MELD Audio Feature Extraction (Optimized)")

print("=" \* 60)

*# Configuration*

audio\_dir = '/content/drive/MyDrive/MELD.Raw/train\_audio'

n\_mfcc = 13

batch\_size = 100  *# Process in batches for better performance*

print(**f**"📁 Audio directory: {audio\_dir}")

print(**f**"📊 Dataset size: {len(train\_df)**:,**} samples")

print(**f**"🎯 MFCC coefficients: {n\_mfcc}")

print(**f**"⚡ Batch size: {batch\_size}")

*# Analyze missing files first*

missing\_dialogues = analyze\_missing\_patterns(train\_df, audio\_dir)

*# Process with optimized method*

print(**f**"\n🚀 Starting optimized processing...")

train\_audio\_features, missing\_count = process\_audio\_optimized(

    train\_df, audio\_dir, n\_mfcc, batch\_size

)

print(**f**"\n📈 Final Results:")

print(**f**"   Features shape: {train\_audio\_features.shape}")

print(**f**"   Expected shape: ({len(train\_df)}, {n\_mfcc})")

print(**f**"   Missing files handled: {missing\_count**:,**}")

*# Save the features*

output\_path = '/content/drive/MyDrive/MELD.Raw/train\_audio\_features.npy'

np.save(output\_path, train\_audio\_features)

print(**f**"💾 Features saved to: {output\_path}")

*# Create a metadata file*

metadata = {

    'total\_samples': len(train\_df),

    'successful\_extractions': len(train\_df) - missing\_count,

    'missing\_files': missing\_count,

    'success\_rate': (len(train\_df) - missing\_count) / len(train\_df) \* 100,

    'n\_mfcc': n\_mfcc,

    'feature\_shape': train\_audio\_features.shape

}

import json

metadata\_path = '/content/drive/MyDrive/MELD.Raw/train\_audio\_features\_metadata.json'

with open(metadata\_path, 'w') as f:

    json.dump(metadata, f, indent=2)

print(**f**"📋 Metadata saved to: {metadata\_path}")

*# Feature statistics*

non\_zero\_samples = np.count\_nonzero(np.sum(train\_audio\_features, axis=1))

print(**f**"\n🔍 Feature Statistics:")

print(**f**"   Non-zero feature samples: {non\_zero\_samples**:,**}")

print(**f**"   Zero-padded samples: {len(train\_df) - non\_zero\_samples**:,**}")

print(**f**"   Mean feature value: {np.mean(train\_audio\_features[train\_audio\_features != 0])**:.4f**}")

print(**f**"   Std feature value: {np.std(train\_audio\_features[train\_audio\_features != 0])**:.4f**}")

print("\n✅ Optimized audio feature extraction complete!")

print("🎯 Ready for emotion classification model training")

4. VIDEO EXTRACTION AND PREPROCESSING

"""## Video Feature Extraction Pipeline

Install Required Libraries

"""

!pip install opencv-python-headless torchvision --quiet

"""Import Dependencies"""

import os

import cv2

import numpy as np

import torch

import torchvision.transforms as transforms

from torchvision.models import resnet50

from tqdm import tqdm

"""Load Pretrained ResNet50 (without classification layer)"""

device = torch.device("cuda" if torch.cuda.is\_available() else "cpu")

resnet = resnet50(pretrained=True)

resnet = torch.nn.Sequential(\*list(resnet.children())[:-1])  *# Remove final classifier*

resnet.to(device).eval()

transform = transforms.Compose([

    transforms.ToPILImage(),

    transforms.Resize((224, 224)),

    transforms.ToTensor(),

    transforms.Normalize(mean=[0.485, 0.456, 0.406],

                         std=[0.229, 0.224, 0.225])

])

"""Frame Sampling & Feature Extraction"""

**def** extract\_frames(video\_path, max\_frames=10):

    cap = cv2.VideoCapture(video\_path)

    total\_frames = int(cap.get(cv2.CAP\_PROP\_FRAME\_COUNT))

    frames = []

    step = max(1, total\_frames // max\_frames)

    for i in range(0, total\_frames, step):

        cap.set(cv2.CAP\_PROP\_POS\_FRAMES, i)

        ret, frame = cap.read()

        if not ret:

            continue

        frame = cv2.cvtColor(frame, cv2.COLOR\_BGR2RGB)

        frames.append(frame)

        if len(frames) >= max\_frames:

            break

    cap.release()

    return frames

**def** extract\_video\_features(video\_path, max\_frames=10):

    frames = extract\_frames(video\_path, max\_frames)

    features = []

    for frame in frames:

        try:

            input\_tensor = transform(frame).unsqueeze(0).to(device)

            with torch.no\_grad():

                embedding = resnet(input\_tensor).squeeze().cpu().numpy()

            features.append(embedding)

        except Exception:

            continue

    return np.mean(features, axis=0) if features else np.zeros(2048)

"""Process All Videos in a Dataset Split"""

**def** process\_video\_split(df, video\_dir, output\_path):

    all\_features = []

    missing = 0

    for \_, row in tqdm(df.iterrows(), total=len(df), desc="Extracting video features"):

        file\_name = **f**'dia{row["Dialogue\_ID"]}\_utt{row["Utterance\_ID"]}.mp4'

        video\_path = os.path.join(video\_dir, file\_name)

        if not os.path.exists(video\_path):

            all\_features.append(np.zeros(2048))

            missing += 1

            continue

        features = extract\_video\_features(video\_path)

        all\_features.append(features)

    all\_features = np.array(all\_features)

    np.save(output\_path, all\_features)

    print(**f**"\n✅ Saved {len(df)} video features to: {output\_path}")

    print(**f**"🚫 Missing files: {missing}")

    print(**f**"📐 Feature shape: {all\_features.shape}")

"""Run on Train Split"""

video\_dir = '/content/drive/MyDrive/MELD.Raw/train/train\_splits'

output\_path = '/content/drive/MyDrive/MELD.Raw/train\_video\_features.npy'

import multiprocessing as mp

from functools import partial

import numpy as np

**def** process\_single\_video(args):

    """Process a single video - designed for multiprocessing"""

    row, video\_dir = args

    file\_name = **f**'dia{row["Dialogue\_ID"]}\_utt{row["Utterance\_ID"]}.mp4'

    video\_path = os.path.join(video\_dir, file\_name)

    if not os.path.exists(video\_path):

        return np.zeros(2048)

    return extract\_video\_features(video\_path)

**def** process\_video\_split\_fast(df, video\_dir, output\_path, num\_workers=4):

    """Multi-threaded video processing"""

*# Prepare arguments for multiprocessing*

    args = [(row, video\_dir) for \_, row in df.iterrows()]

*# Process in parallel*

    with mp.Pool(num\_workers) as pool:

        all\_features = list(tqdm(

            pool.imap(process\_single\_video, args),

            total=len(df),

            desc="Extracting video features (parallel)"

        ))

    all\_features = np.array(all\_features)

    np.save(output\_path, all\_features)

    print(**f**"✅ Saved {len(df)} video features to: {output\_path}")

    print(**f**"📐 Feature shape: {all\_features.shape}")

*# Usage*

num\_cores = mp.cpu\_count()

print(**f**"Using {num\_cores} CPU cores")

process\_video\_split\_fast(train\_df, video\_dir, output\_path, num\_workers=num\_cores)

5. Simple MLP model as baseline

"""Train a Simple Multimodal Model (MLP)"""

from sklearn.neural\_network import MLPClassifier

from sklearn.metrics import classification\_report

model = MLPClassifier(hidden\_layer\_sizes=(512, 256), max\_iter=100, random\_state=42)

model.fit(X\_train, y\_train\_encoded)

"""Evaluate on Validation Set"""

*# Evaluate*

y\_pred = model.predict(X\_train\_dev)

print(classification\_report(y\_dev\_encoded, y\_pred, target\_names=le.classes\_))

"""Test and Save Final Results"""

y\_pred\_test = model.predict(X\_train\_test)

print(classification\_report(y\_test\_encoded, y\_pred\_test, target\_names=le.classes\_))

6. Defining and fine-tuning final model+ confidence thresholding and ambivalence flgging mechanism

"""Define the Multimodal MLP"""

import torch.nn as nn

import torch.nn.functional as F

**class** MultimodalMLP(nn.Module):

**def** \_\_init\_\_(self, input\_dim, num\_classes):

        super(MultimodalMLP, self).\_\_init\_\_()

        self.fc1 = nn.Linear(input\_dim, 1024)

        self.bn1 = nn.BatchNorm1d(1024)

        self.fc2 = nn.Linear(1024, 512)

        self.bn2 = nn.BatchNorm1d(512)

        self.fc3 = nn.Linear(512, 256)

        self.bn3 = nn.BatchNorm1d(256)

        self.fc4 = nn.Linear(256, 128)

        self.dropout = nn.Dropout(0.4)

        self.out = nn.Linear(128, num\_classes)

**def** forward(self, x):

        x = self.bn1(F.relu(self.fc1(x)))

        x = self.bn2(F.relu(self.fc2(x)))

        x = self.bn3(F.relu(self.fc3(x)))

        x = self.dropout(F.relu(self.fc4(x)))

        return self.out(x)

"""Set Up Training with Class Weights & LR Scheduler"""

*# Class weights*

from sklearn.utils.class\_weight import compute\_class\_weight

import numpy as np

*# Define the device*

device = torch.device("cuda" if torch.cuda.is\_available() else "cpu")

print(**f**"Using device: {device}")

*# Class weights*

from sklearn.utils.class\_weight import compute\_class\_weight

class\_weights = compute\_class\_weight('balanced', classes=np.unique(y\_train\_encoded), y=y\_train\_encoded)

weights\_tensor = torch.tensor(class\_weights, dtype=torch.float32).to(device)

*# Model setup*

model = MultimodalMLP(input\_dim=X\_train.shape[1], num\_classes=len(le.classes\_)).to(device)

optimizer = torch.optim.Adam(model.parameters(), lr=1e-4)

scheduler = torch.optim.lr\_scheduler.StepLR(optimizer, step\_size=3, gamma=0.5)

loss\_fn = nn.CrossEntropyLoss(weight=weights\_tensor)

"""Train Model with Validation Accuracy Tracking"""

**def** train\_model(model, train\_loader, dev\_loader, optimizer, scheduler, loss\_fn, epochs=15):

    best\_acc = 0.0

    for epoch in range(epochs):

        model.train()

        total\_loss = 0

        for X\_batch, y\_batch in train\_loader:

            X\_batch, y\_batch = X\_batch.to(device), y\_batch.to(device)

            optimizer.zero\_grad()

            outputs = model(X\_batch)

            loss = loss\_fn(outputs, y\_batch)

            loss.backward()

            optimizer.step()

            total\_loss += loss.item()

        scheduler.step()

*# Validation*

        model.eval()

        correct, total = 0, 0

        with torch.no\_grad():

            for X\_val, y\_val in dev\_loader:

                X\_val, y\_val = X\_val.to(device), y\_val.to(device)

                preds = torch.argmax(model(X\_val), dim=1)

                correct += (preds == y\_val).sum().item()

                total += y\_val.size(0)

        val\_acc = correct / total

        print(**f**"Epoch {epoch+1} | Train Loss: {total\_loss**:.4f**} | Dev Acc: {val\_acc**:.4f**}")

*# Save best model*

        if val\_acc > best\_acc:

            best\_acc = val\_acc

            torch.save(model.state\_dict(), "best\_multimodal\_model.pt") train\_model(model, train\_loader, dev\_loader, optimizer, scheduler, loss\_fn, epochs=15)

"""Evaluation Function"""

from sklearn.metrics import classification\_report

import numpy as np

**def** evaluate\_model(model, X, y\_true, threshold=0.6, ambivalence\_delta=0.1):

    model.eval()

    with torch.no\_grad():

        inputs = torch.tensor(X, dtype=torch.float32).to(device)

        logits = model(inputs)

        probs = torch.softmax(logits, dim=1).cpu().numpy()

        preds = np.argmax(probs, axis=1)

*# Confidence thresholding*

        confidences = probs.max(axis=1)

        low\_conf\_mask = confidences < threshold

*# Ambivalence detection*

        sorted\_probs = np.sort(probs, axis=1)

        ambiv\_diff = sorted\_probs[:, -1] - sorted\_probs[:, -2]

        ambiv\_mask = ambiv\_diff < ambivalence\_delta

        print("🎯 Overall Performance:")

        print(classification\_report(y\_true, preds, target\_names=le.classes\_))

        print(**f**"\n⚠️ Low-Confidence Predictions: {np.sum(low\_conf\_mask)} / {len(y\_true)}")

        print(classification\_report(np.array(y\_true)[low\_conf\_mask], preds[low\_conf\_mask], target\_names=le.classes\_))

        print(**f**"\n🌀 Ambivalent Predictions: {np.sum(ambiv\_mask)} / {len(y\_true)}")

        print(classification\_report(np.array(y\_true)[ambiv\_mask], preds[ambiv\_mask], target\_names=le.classes\_))

        return probs, preds, confidences, low\_conf\_mask, ambiv\_mask

"""Load Best Model & Evaluate"""

*# Load best model*

best\_model = MultimodalMLP(input\_dim=X\_train.shape[1], num\_classes=len(le.classes\_)).to(device)

best\_model.load\_state\_dict(torch.load("best\_multimodal\_model.pt"))

best\_model.eval()

*# Evaluate on dev*

print("\n📊 Evaluation on Dev Set")

evaluate\_model(best\_model, X\_dev\_scaled, y\_dev\_encoded)

*# Evaluate on test*

print("\n📊 Evaluation on Test Set")

evaluate\_model(best\_model, X\_test\_scaled, y\_test\_encoded)

7. EMOJI MAPPING, CONFIDENCE AND AMBIVALENCE FLAGGING DISPLAY, DEMO SET-UP

**def** show\_demo(index, df, X\_scaled, video\_dir, model):

    row = df.iloc[index]

    true\_emotion = row['Emotion']

    dialogue\_id = row['Dialogue\_ID']

    utterance\_id = row['Utterance\_ID']

    text = row['Utterance']

    source\_video\_dir = '/content/drive/MyDrive/MELD.Raw/test/output\_repeated\_splits\_test'

    video\_filename = **f**"dia{dialogue\_id}\_utt{utterance\_id}.mp4"

    video\_path = os.path.join(source\_video\_dir, video\_filename)

*# Display video*

    print(**f**"🎞 Showing video: {video\_filename}")

*# Added embed=True to correctly display the video from a local file path*

    display(Video(video\_path, width=320, embed=True))

*# Model inference*

*# Ensure 'device' variable is defined (it's in Global Variables)*

    x = torch.tensor(X\_scaled[index], dtype=torch.float32).unsqueeze(0).to(device)

    best\_model.eval() *# Set model to evaluation mode*

    with torch.no\_grad(): *# Disable gradient calculation*

        logits = model(x)

    probs = torch.softmax(logits, dim=1).squeeze().cpu().numpy()

*# Ensure 'le' (LabelEncoder) variable is defined (it's in Global Variables)*

    top1 = np.argmax(probs)

*# Ensure there are at least 2 classes to get top2*

    if len(le.classes\_) > 1:

        top2 = np.argsort(probs)[-2]

        ambiv\_margin = probs[top1] - probs[top2]

    else:

        top2 = -1 *# No second best if only one class*

        ambiv\_margin = 0.0 *# No ambivalence if only one class*

    confidence = probs[top1]

    pred\_emotion = le.classes\_[top1]

*# Flag logic*

    low\_conf\_flag = confidence < 0.6

*# Check ambivalence only if there's more than one class*

    ambiv\_flag = False

    if len(le.classes\_) > 1:

        ambiv\_flag = ambiv\_margin < 0.1

*# Display info*

    print(**f**"\n🗣 Utterance: {text}")

    print(**f**"🎯 Predicted Emotion: {pred\_emotion}")

    print(**f**"🔒 Confidence: {confidence**:.3f**}")

*# Only print ambivalence if applicable*

    if len(le.classes\_) > 1:

        print(**f**"🌀 Ambivalence Margin: {ambiv\_margin**:.3f**}")

    print(**f**"✅ True Emotion: {true\_emotion}")

*# Flags*

    if low\_conf\_flag:

        print("⚠️ Low confidence prediction")

    if ambiv\_flag:

        print("🌀 Ambivalent (Top-2 emotions too close)")

*# Show probability table*

    prob\_df = pd.DataFrame({

        'Emotion': le.classes\_,

        'Probability': [**f**"{p**:.2%**}" for p in probs]

    }).sort\_values(by='Probability', ascending=False)

    display(prob\_df.reset\_index(drop=True))

*# Human-readable emoji summary*

    emoji\_map = {

        'anger': '😠', 'disgust': '🤢', 'fear': '😨',

        'joy': '😄', 'neutral': '😐', 'sadness': '😢', 'surprise': '😲'

    }

    print("\n📋 Summary:")

    print(**f**"🎯 Prediction: {emoji\_map.get(pred\_emotion, '')} {pred\_emotion} ({confidence**:.1%**} confidence)")

    print(**f**"✅ Ground Truth: {emoji\_map.get(true\_emotion, '')} {true\_emotion}")

    if low\_conf\_flag:

        print("⚠️ Flagged: Low confidence")

    if ambiv\_flag:

        print("🌀 Flagged: Ambivalent (emotion overlap)")