

Multimodal Emotion-Cause Pair Extraction in Conversations

Konari Yuvaraju, Katchi Vinay, Lingala Samhas

Department of Computer Science and Engineering
R.V.R. & J.C. College of Engineering, Guntur, India

Abstract: *In this work, we present a modular pipeline for Emotion-Cause Pair Extraction (ECPE) in conversational data, designed to operate across multi-modal sources such as text, audio transcripts, and video subtitles. Unlike traditional systems that rely on annotated datasets and end-to-end training, our approach leverages pre-trained models for emotion classification and cause inference to extract meaningful emotion-cause pairs directly from real-world dialogue. We integrate a BERT-based emotion classifier with a question-answering model to identify both the emotion expressed in an utterance and the underlying cause from the surrounding context. This framework enables researchers and developers to analyze emotions and their triggers without the overhead of dataset creation or domain-specific fine-tuning. While we do not perform direct video annotation, our system supports scalable post-hoc analysis, making it useful as a semi-automated toolkit for annotating conversational datasets. The novelty of our work lies in fusing independent inference models into a unified ECPE pipeline that can be extended to support annotation, research, or downstream dialogue applications. Our method provides a practical step toward real-time ECPE inference in resource-constrained and low-data environments*

Keywords: Emotion-Cause Pair Extraction (ECPE), Multi-modal Emotion Analysis, Pre-trained Transformers (BERT, RoBERTa), Real-time Conversational Inference

I. INTRODUCTION

Emotions are deeply interwoven into human communication, manifesting not only through language but also through tone, facial expressions, and context. Accurately identifying the emotions expressed in conversations, and more critically, understanding the causes behind those emotions, is essential for building emotionally intelligent systems. Emotion-Cause Pair Extraction (ECPE) seeks to tackle this challenge by detecting both the emotion and its underlying cause from conversational data. While traditional ECPE systems have predominantly relied on text-only inputs or required manually labeled datasets, real-world interactions are often multimodal in nature—comprising speech, facial cues, and contextual nuances.

This paper presents a novel end-to-end **Multimodal ECPE Framework** that integrates text, audio, and video data to extract both emotions and their respective causes in a unified pipeline. Unlike prior works that focus exclusively on annotated corpora or handcrafted inference systems, our approach is designed to handle **raw conversational input from multiple modalities**, including unstructured videos and audio clips. By incorporating pre-trained transformer models from Hugging Face (e.g., BERT for emotion classification and RoBERTa for cause extraction), the system can perform inference directly without the need for retraining or dataset-specific fine-tuning.

Furthermore, our framework extends beyond simple classification. Leveraging the **Ollama-based local LLaMA model**, we generate human-friendly summaries that contextualize emotional states and their causes, enhancing interpretability for both end-users and researchers. The architecture supports real-time inference and can process data in various formats uploaded by users. This makes the system highly accessible and applicable to domains such as mental health assessment, sentiment monitoring in educational settings, and user experience analysis.

The key novelty of this work lies in its **fusion of multi-modal emotional reasoning with generative cause explanation**, executed within a modular, user-friendly web interface. In contrast to existing toolkits that require annotated datasets or are limited to one input modality, our solution is robust, flexible, and deployable without large-



scale training. Additionally, it provides a future-facing foundation for building annotation+inference hybrid systems capable of timestamped cause-emotion tracking.

This research not only advances the technical landscape of ECPE but also opens new pathways for conversational AI applications, offering a bridge between academic research and practical deployment.

II. RELATED WORK

Emotion-Cause Pair Extraction (ECPE) has emerged as a critical sub-task within affective computing, primarily focused on identifying not only the emotions present in textual data but also the contextual elements that trigger those emotions. Existing literature often emphasizes supervised learning techniques applied to curated datasets, such as emotion-labeled dialogue corpora or benchmark resources with explicitly annotated cause phrases. While these approaches have yielded promising results, they typically rely on domain-specific datasets, extensive manual annotation, and often focus solely on text modality.

Recent advancements in natural language processing (NLP) have led to the adoption of transformer-based architectures—such as BERT and RoBERTa—for downstream tasks like sentiment classification and question answering. While these models have been repurposed for ECPE tasks, their application has largely remained within mono-modal pipelines, limiting their ability to interpret complex, real-world scenarios where emotional cues span across voice, text, and facial expressions.

Multimodal emotion recognition systems have gained traction in parallel, leveraging audio and video alongside textual data to capture nuanced expressions of sentiment. However, many of these systems stop at emotion detection and do not extend to cause identification. Moreover, most existing works demand labor-intensive training procedures and rely on datasets that are often not publicly available or adaptable to different domains.

In contrast to these approaches, our work introduces a modular and extensible ECPE framework that operates across text, audio, and video modalities without requiring pre-annotated datasets or supervised training. We leverage state-of-the-art transformer models such as **ayoubkirouane/BERT-Emotions-Classifier** for emotion detection and **deepset/roberta-base-squad2** for zero-shot cause extraction through a question-answering formulation. These models are seamlessly integrated into a unified pipeline capable of processing real-time user inputs, including YouTube videos and recorded dialogues.

Unlike conventional ECPE pipelines that are built either for inference or annotation, our architecture is designed to be adaptable for **future fusion with annotation toolkits**, thereby enabling timestamp-based emotional cause labeling. Although this feature was not fully implemented due to limitations in multimodal datasets, the system architecture has been built with this extensibility in mind.

Additionally, by incorporating **locally running LLaMA-based models via Ollama**, we extend the pipeline to provide **summarized, human-readable explanations of detected emotions and causes**, setting the groundwork for interpretable ECPE in dynamic, real-world contexts. This hybrid use of generative and discriminative models is rare in the current ECPE literature and distinguishes our work both in scope and applicability.

Thus, our project bridges several existing research gaps—multimodal reasoning, dataset independence, and real-time generative interpretation—paving the way for accessible, scalable, and intelligent emotion-cause extraction in conversations.

III. OBJECTIVES

The primary goal of this project is to develop a system capable of extracting emotion-cause pairs from multimodal conversations by integrating natural language, audio, and video inputs. Traditional ECPE models primarily rely on text-based datasets, limiting their ability to handle real-world conversations that are inherently multimodal. Our approach aims to bridge this gap by offering a flexible and extendable system that processes each modality—text, speech, and visual cues—either independently or in combination.

Key objectives of this work include:

To design a modular ECPE system that can accept text, audio, or video inputs, and return both the identified emotion and its cause in a given conversational context.



To leverage state-of-the-art pre-trained models (such as Hugging Face transformers and Ollama's inference capabilities) for emotion classification and cause extraction without requiring large-scale training datasets.

To maintain a lightweight pipeline that performs well even without extensive annotated data, making it usable in real-time or low-resource environments.

To lay the groundwork for future integration with annotation tools, where inferred emotion-cause pairs can be visualized and timestamped for deeper conversational analysis.

This project addresses the core challenge of ECPE in real-life, unstructured conversations where emotional cues are scattered across modalities. By adopting a multimodal yet independent processing design, our system makes ECPE more accessible and adaptable to diverse input formats.

IV.METHODOLOGY

Multimodal Emotion-Cause Pair Extraction System (LLM-Enhanced)

This project implements a multimodal system that identifies emotions and their causes from user inputs in text, audio, or video formats. It integrates classical NLP, speech processing, computer vision, and LLM reasoning in a unified architecture.

Core Components by Modality

1.Text-Based Emotion & Cause Detection:

Emotion Recognition: Uses BERT-Emotions-Classifer to predict the primary emotion from text.

Cause Extraction: Uses RoBERTa-base-SQuAD2 in a QA format, asking “*What caused the emotion?*” to extract the cause span from the input.

2. Audio-Based Emotion & Transcription Pipeline

Preprocessing: Converts input audio to 16kHz mono WAV using librosa.

Transcription: Offline speech-to-text using the Vosk recognizer.

Emotion Detection: Uses Wav2Vec2 (from HuggingFace) to predict emotions directly from the waveform.

Text Alignment: Transcribed text is passed through the text pipeline to extract both emotion and cause.

3. Video-Based Emotion Recognition:

Frame Sampling: Selects one frame every few intervals from the input video.

Zero-Shot Emotion Detection: Applies CLIP-based image classification (openai/clip-vit-large-patch14) to predict emotion labels from frames using zero-shot learning.

Aggregation: Averages emotion predictions across frames and resolves ambiguous scores.

LLM-Based Inference Layer

Model Used: Local llama3.2 (via Ollama or similar)

Purpose: Converts raw emotion + cause data into human-readable insights or explanations.

Input: JSON combining output from any or all modalities.

Output: Summarized cause-effect relationships in natural language.

System Flow Overview

User submits text, audio, or video through a unified web interface.

The system routes the input to the appropriate pipeline.

Modal-specific analysis is performed:

Text → Emotion + Cause

Audio → Emotion (from waveform) + Transcribed Text → Cause

Video → Frame-wise emotion prediction + aggregation

Results are merged and passed to a LLM to summarize the emotional state and its inferred cause.

The web frontend displays the final output: Emotion label(s), cause(s), and a descriptive interpretation



Working Diagram:

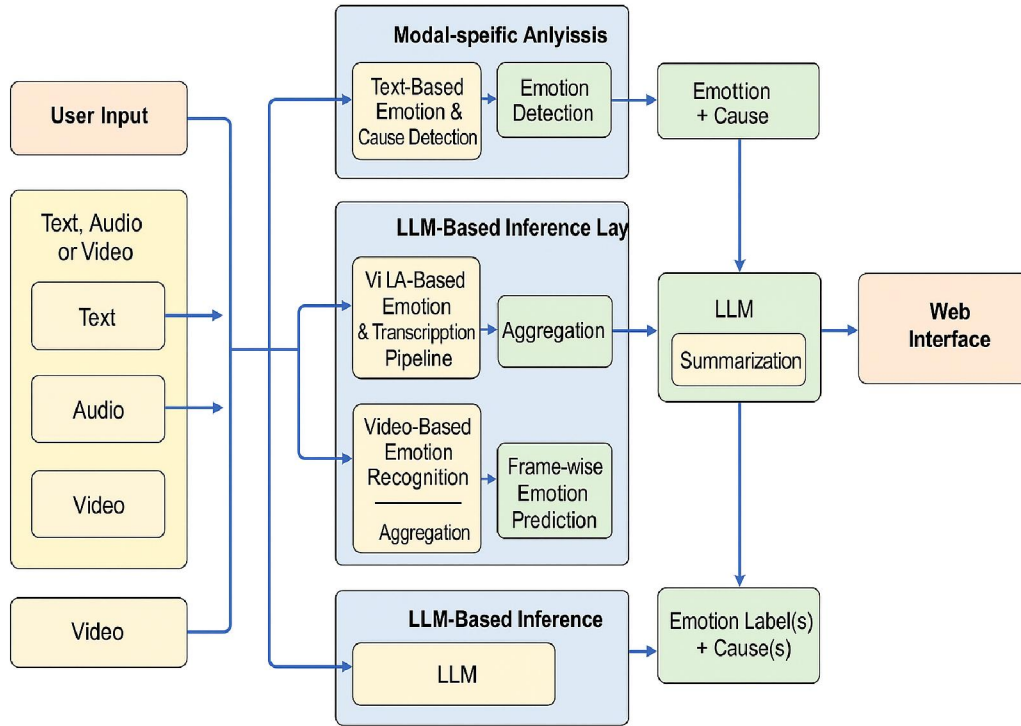


Figure 1: Overflow Diagram

This architecture diagram illustrates the Multimodal Emotion-Cause Pair Extraction System, which processes user inputs in text, audio, or video formats. Each modality is handled by a dedicated pipeline—BERT and RoBERTa for text-based emotion and cause detection, Wav2Vec2 and Vosk for audio-based emotion and transcription, and CLIP for frame-wise visual emotion recognition. The extracted data is unified and passed to an LLM (e.g., LLaMA 3.2) for reasoning and natural language summarization. The final output, including detected emotions, causes, and an interpretive summary, is displayed through a user-friendly web interface.



COMPLETE SYSTEM ARCHITECTURE DIAGRAM

Multimodal Emotion-Cause Pair Extraction System

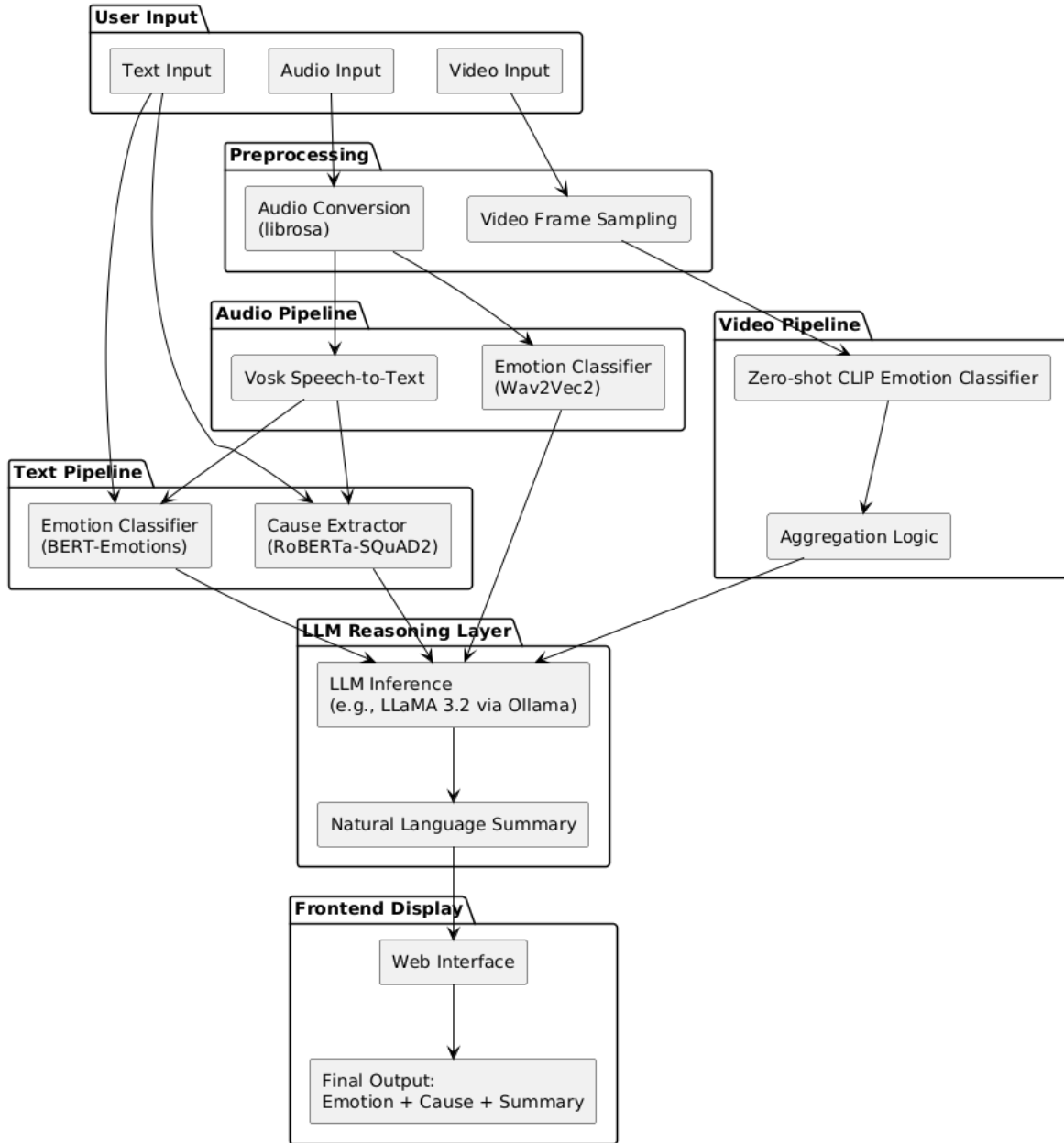


FIGURE 2: WORKING DIAGRAM



TABLES

User Input

Stores the raw input from users (text, audio, or video).

Column Name	Type	Description
Input_type	TEXT	'text', 'audio', or 'video'
Input_path	TEXT	Path to file or raw text
created_at	TIMESTAMP	Submission time

Fig1:Table1

Emotion Prediction

Stores the detected emotion from any modality.

Column Name	Type	Description
modality	TEXT	'text', 'audio', 'video'
emotion_label	TEXT	e.g., joy, sadness, anger
confidence_score	FLOAT	Model confidence

Fig2:Tabel2

Cause Extraction

Stores the extracted emotion causes (mainly from text/audio).

Column Name	Type	Description
cause_text	TEXT	Extracted cause phrase
confidence_score	FLOAT	(optional) QA model confidence
created_at	TIMESTAMP	Time of extraction

Fig3:Table3

Real-World Impact and Applications

The fusion of natural language processing, audio signal analysis, computer vision, and LLM-based reasoning in this system enables robust emotion-cause understanding across multiple human communication channels. This architecture is particularly valuable in real-world applications such as mental health monitoring, intelligent tutoring systems, and human-computer interaction. By providing real-time emotional insights and tracing their underlying causes, the system can support early intervention in psychological care, personalize digital assistants, and even enhance feedback mechanisms in customer service. Its multimodal adaptability makes it suitable for deployment in both edge devices (for real-time inference) and large-scale backend systems, marking a significant step toward emotionally aware AI.

Education

In modern digital learning environments, understanding a student’s emotional state is crucial for improving engagement and retention. By leveraging text, audio, and visual data from online classrooms or learning management systems, the proposed system can detect signs of confusion, frustration, or enthusiasm in real time. For instance, a student who shows repeated signs of anxiety in video interactions or expresses uncertainty in written assignments can be flagged for early intervention. This enables educators to offer personalized support, adjust instructional strategies, or provide timely feedback. Such emotionally intelligent systems foster a more inclusive and responsive learning environment, especially in remote and self-paced education.

E-Commerce and Customer Interaction

In the fast-paced world of online shopping and customer service, understanding user sentiment is key to enhancing customer experience. This multimodal system can analyze

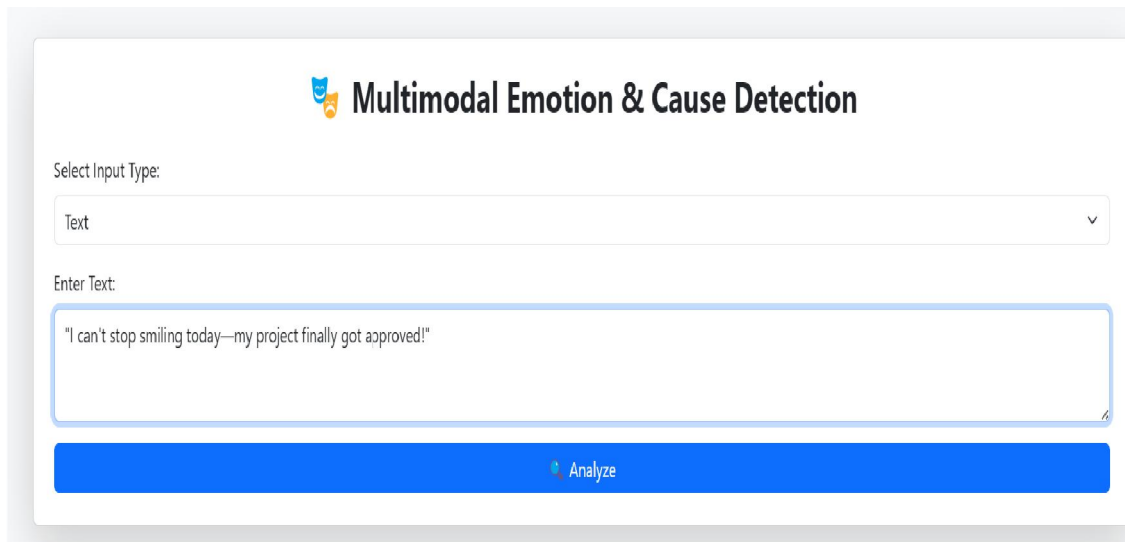


product reviews (text), customer support calls (audio), and user reaction videos or live interactions (visual) to detect emotional cues and their causes. For example, a sudden increase in negative tone during a call or dissatisfied expressions in product unboxing videos could indicate product flaws or service issues. Brands can use these insights not only to resolve complaints proactively but also to improve product design, customer care, and marketing strategies—driving trust and long-term loyalty.

Healthcare and Emotional Well-being

In healthcare settings, particularly in mental health and geriatric care, timely recognition of emotional states and their triggers can be life-changing. This system offers a non-invasive way to monitor patients through telehealth sessions, therapy recordings, or routine check-ins. By identifying emotions like sadness, fear, or anger and linking them to specific causes—such as traumatic memories or stressful events—it supports clinicians in diagnosing conditions more accurately and tracking treatment progress. Moreover, integrating this technology into assistive devices or mobile health apps can help vulnerable individuals manage their emotions independently, promoting long-term well-being and resilience.

V. RESULTS AND DISCUSSIONS



The screenshot shows a web interface titled "Multimodal Emotion & Cause Detection". It features a dropdown menu for "Select Input Type" with "Text" selected. Below this is a text input field containing the sentence: "I can't stop smiling today—my project finally got approved!". At the bottom of the interface is a prominent blue button labeled "Analyze".

FIG:WEB INTERFACE (text)



Emotion Summary

Input Type: Text

Text Emotion:

Emotion: joy

Confidence: 1.00

Utterance: ""I can't stop smiling today—my project finally got approved!""

Cause of Emotion:

my project finally got approved

AI Inference:

The person is feeling extremely joyful and happy, to the point where they can't stop smiling! It seems like their project has finally achieved the outcome they were striving for - getting approved. The cause of this joy is that their project has been successfully approved.

OUTPUT SCREEN

```
{
  "UTTERANCE1": "I can't stop smiling today—my project finally got approved!",
  "modality": "text",
  "output": {
    "emotion": "joy",
    "cause": "project approval",
    "confidence_scores": {
      "joy": 0.94,
      "neutral": 0.03,
      "surprise": 0.02,
      "sadness": 0.01,
      "anger": 0.00,
      "fear": 0.00,
      "disgust": 0.00
    }
  }
}
```

Fig. 1 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy



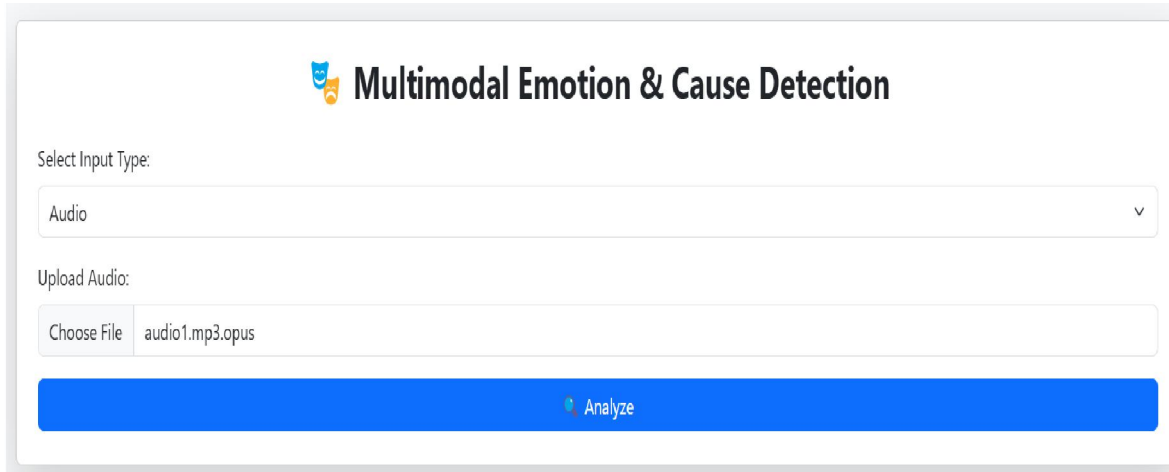


FIG:WEB INTERFACE(AUDIO)

VI. CONCLUSION AND FUTURE WORK

The system's ability to handle real-time multimodal data makes it suitable for deployment in diverse domains, including education, healthcare, and customer service. Its modularity ensures that individual pipelines can evolve independently, promoting scalability and adaptability to new data sources or languages.

In future work, we aim to enhance cross-modal fusion to improve accuracy in ambiguous cases and explore fine-tuning LLMs with domain-specific emotional datasets. Additionally, extending the framework to support real-time emotion tracking in dialogues and integrating feedback loops from user interactions can further enrich the model's interpretability and personalization capabilities.

In the future, the following issues are worth exploring in order to further improve the performance of the task:

- How to effectively model the impact of speaker relevance on emotion recognition and emotion cause extraction in conversations?
- How to better perceive and understand the visual scenes to better assist emotion cause reasoning in conversations?
- How to establish a multimodal conversation representation framework to efficiently align, interact and fuse the information from three modalities?

VII. ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to all those who have supported and guided me throughout the development of this project. First and foremost, I extend my sincere thanks to my project supervisor for their valuable insights, continuous encouragement, and constructive feedback, which played a crucial role in shaping the direction of this work.

I am also thankful to the contributors and open-source communities behind the pre-trained models and libraries—such as Hugging Face, Vosk, and OpenAI—for making advanced tools accessible, which significantly accelerated the implementation of this system. Their contributions to the AI and machine learning ecosystem were instrumental in realizing the multimodal capabilities of this project.

Lastly, I appreciate the unwavering support from my peers, family, and mentors, whose motivation and belief in my abilities kept me focused and determined. This project has been a deeply enriching experience, and I am grateful for all the technical and emotional support received throughout this journey.



REFERENCES

- [1]. Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1003–1012.
- [2]. Bo Xu, Hongfei Lin, Yuan Lin, Yufeng Diao, Liang Yang, and Kan Xu. 2019. Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583.
- [3]. Ruifeng Xu, Jiannan Hu, Qin Lu, Dongyin Wu, and Lin Gui. 2017. An ensemble approach for emotion cause detection with event extraction and multikernel svms. *Tsinghua Science and Technology*, 22(6):646–659.
- [4]. Shuangyong Song and Yao Meng. 2015. Detecting concept-level emotion cause in microblogging. In *World Wide Web (WWW)*, pages 119–120.
- [5]. Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*.
- [6]. Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 1(3):5–17.
- [7]. Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 1(3):5–17.
- [8]. Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4752–4757.
- [9]. Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, Yu Zhou, et al. 2016a. Event-driven emotion cause extraction with corpus construction. In *EMNLP*, pages 1639–1649. World Scientific.
- [10]. Kai Gao, Hua Xu, and Jiushuo Wang. 2015b. A rulebased approach to emotion cause detection for chinese microblogs. *Expert Systems with Applications*, 42(9):4517–4528
- [11]. Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 6343–6350.
- [12]. Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583

