Emotion and Sentiment Evolution

Sanyukta Deshmukh[1], Advait Ugavekar[2], Aary Upare[3], Swanand Uttarwar[4], Vishal Wale[5]
[1,2,3,4,5]*Vishwakarma Institute of Technology, Pune*

***Abstract*—The conversation agents grow rapidly creating many limitations many agents are not that connective to the emotional nature of the humans, which affect the interaction between human and machine. In this paper we are going to study about the Artificial Intelligence - AI for affective computation. We are going to keep track of the progress of emotion detection from basic text to advanced reasoning. We are going to start with basic text-based method of conversation used in most ML models such as Naïve Bayes and Support Vector Machines (SVM), and we are going to compare them with deep learning neural networks which achieves high accuracy (for example, an F1-score of 0.95 for sadness detection). Then we are going to move toward context-based Emotion Recognition in Conversation (ERC), also going to focus on some pre-trained language models like BERT which are fine-tuned for including dialogue structure and speaker information. Along with this we are going to use the approach of Multimodal Emotion Recognition (MER), using which we can combine text, audio, and some visuals to reduce the "heterogeneity gap," with the help of both supervised and unsupervised learning methods. At the end, we are going to discuss about the final stage of emotion evolution AI systems that understand casualty in emotions. In this we have included models which help us to identify the cause of emotion generation to generate the response which can show empathy, along with this we are trying to use Retrieval-Augmented Generation (RAG) frameworks such as "Cause Motion," which keep the track long-sequence causal patterns in multimodal conversations. Overall, this research focuses on the change from simple emotion classification to a more complete, cause based understanding of human emotions.**

***Index Terms*— Emotion, Sentiment Evolution, Artificial Intelligence (AI), Natural Language Processing (NLP), Affective Computation, Detection of Emotions, Machine Learning, Neural Networks, Human-Machine Interaction, Emotion Recognition in Conversation (ERC), Multimodal Emotion Recognition (MER), Cross-Modal Fusion, Emotional Causality, Retrieval-Augmented Generation (RAG), Empathetic Response Generation**

I. INTRODUCTION

Emotions are an essential part of human life and play a main role in both verbal and nonverbal communication. Detecting and interpreting the emotions accurately is essential for meaningful and emotional interaction. In recent years, conversational agents like chatbots, virtual assistants, and voice-based interfaces have become common in customer support, telemedicine, and e-learning. However, even though modern large language models (LLMs) can generate fluent and natural dialogue, still many of these systems are emotionally unaware, often responding in a uniform and common to all nature.

This creates a major problem or say lack of human and machine interaction. Integrating automatic emotion detection can provide the ability to conversational agents to recognize a user's emotions and generate the response according to that. This can be a big step toward building systems that generate more empathetic, emotionally intellectual and human-like model. The use of such technology is broad and significant. By improving normal chatbot communication and analysing user opinions to support sensitive domains such as mental health counselling and education.

This paper is about "Emotion and Sentiment Evolution" within Artificial Intelligence (AI), and focusing on how emotion recognition became advanced in time. The development in this field can be seen through four stages,

- Foundational Text Classification: In the first stage, we analyze identification of emotion in short text segments like posts and comments using old machine learning models like LR, NB, SVM and then deep learning models.
- Context-Aware Conversational AI (ERC): In the second stage we are going to identify emotions with context as well. In this stage we consider speaker roles, dialogue structure for better

understanding and changes in emotions.

- Multimodal Emotion Recognition (MER): Then we saw the problem that text alone was not able to cover all the emotion. In this stage we are going to talk about models that combine text, speech, and visuals to recognize emotional states more accurately.
- Emotional Causality: The most current stage goes beyond identifying which emotion is present its focus is to understand why the emotion occurs. At this level, systems find why emotions happen and track long-term feeling patterns in talks."

Overall, this study looks at the technology growing through these four stages, reviewing basic and advanced models, and describes how AI is getting near to true emotional understanding.

II. LITERATURE REVIEW

Stage 1: Foundational Text-Based Emotion Detection

His first big problem for AI in affective computing was finding emotions from one still text sample like a post or remark, online message too.As detailed given by Machová et al. [1], early research tested different computational methods for this issue.

The earliest technique was the lexicon-based method, which depends on building a strong lexicon a list of words marked with their emotional polarity or level of link to a certain emotion. The system then predicts the chance of a specific emotion that is based on the words found in the text.

This stage was followed by traditional machine learning (ML) approaches, which were considered "good and accurate enough" for simple text emotion tasks. Popular models are the probabilistic Naïve Bayes (NB) classifier and Support Vector Machine (SVM) were trained using labeled text datasets to identify emotions. But over time, these older methods have become less relevant as the field moved toward deep learning techniques.

Deep learning approaches made emotion detection far more accurate, as they can automatically learn complex semantic patterns. Specifically, Convolutional Neural Networks (CNNs) are applied to extract important features, while Recurrent Neural Networks (RNNs) especially Long Short-Term Memory (LSTM) models handle the ordered structure of text. Machová et al. noted that their neural network system, which combined Conv1D with LSTM, achieved outstanding results in multi-class emotion recognition across six categories (joy, sadness, anger, fear, love, and surprise), reaching an F1-score of 0.95 for sadness.[1]

Stage 2: Evolving to Conversational AI (ERC)

While foundational models execute very well for isolated texts, they fall short in conversational settings. Emotion may keep changing during a conversation and it relies on the speaker's context and past interactions. This led to the development of Emotion Recognition in Conversation (ERC), which studies emotions across dialogues by taking three main aspects into account:

1. The query or main text (the target message to classify)
2. The past context (surrounding utterances)
3. Structure of the dialogue (speaker identity and interaction patterns)

Earlier ERC studies worked by pre-training language models such as RoBERTa which was used to extract features that did not depend on context. After that another classifier was used for example, a Gated Recurrent Unit (GRU) or a Graph Convolutional Network (GCN) which modified the dialogue context using those predictions.

However, Qin et al. [2] said that this method only helped in small gains. Their results demonstrated that a simple baseline model (RoBERTa-large combined with an MLP classifier) reached 63.39% accuracy on the MELD dataset, which was almost the same as more advanced systems like DialogRNN and DialogGCN. The main problem was that features generated by pretrained models were abstract and we needed to keep feature extraction and context learning separate which affected the model.

To address this, researchers introduced a fine-tuning framework in which query, context, and dialogue information are directly integrated during the model's fine-tuning process. The BERT-ERC model uses this approach by applying the following methods.

- Suggestive Text: Uses a new format which includes nearby utterances, speaker;s name a token which indicates the target emotion.
- Fine-Grained Classification Module: Adds a classifier that creates position-aware features without altering how the query flows through the system

- Two-Stage Training: It uses a setup where a model receives contextual emotion knowledge from a different model's predictions.

This Automated Approach Achieved State-Of-The-Art Result [2] (Mer)

The Cross-Modal RoBERTa (CM-RoBERTa) framework, introduced by Luo et al. [3], is specifically built to combine spoken audio with its matching text transcripts.

Although ERC models are effective at capturing contextual information, they continue to depend solely on textual data. In reality, human emotions are conveyed through more than just words "through multiple channels voice, facial expressions, and gestures making emotion inherently multimodal. To achieve richer emotional understanding, research progressed into Multimodal Emotion Recognition (MER), focusing on fusing data from different modalities.

3.1 Supervised Cross-Modal Fusion: Text and Speech

One challenge in MER is to connect information from different type of data. Luo et al. introduced Cross-Modal RoBERTa (CM-RoBERTa), a supervised machine learning model that is used to combine spoken audio and text.

- Text: Taken from RoBERTa encoder.
- Audio Features: created by combining 6,552 handcrafted features (pitch, loudness, etc.) through openSMILE and high-level "bottleneck embeddings" from the pre-trained OpenL3 model.

The main part of CM-RoBERTa is the Self- and Cross-Attention (SCA) unit, which learns from both intra-modal and inter-modal relationships.

- Self-Attention learns patterns from each modality separately.
- Cross-Attention learns how one modality can explain another like text and audio

For example, the words "Excuse Me" can mean many emotions. But when the audio has a tone or sounds like anger, the model understands it as angry. With mid-level fusion and residual connections, the model can capture long-range patterns and learn subtle relationships between text and audio

3.2 Unsupervised Multimodal Fusion: Text, Audio, and Vision

A new approach is the Modality-Pairwise Unsupervised Contrastive Loss framework introduced by Franceschini et al. [4]. Supervised MER models are powerful but require large labelled datasets, which are costly and time consuming to create because of collection, structuring and annotating. To overcome this, researchers developed a new method which is unsupervised learning methods, such as the Modality-Pairwise Unsupervised Contrastive Loss framework.

This approach learns from unlabelled text, audio, and visual data by comparing modality pairs (e.g., text–audio, audio–video).

- Positive pairs (same sample) are pulled closer in the embedding space.
- Negative pairs (different samples) are pushed apart.

This instance discrimination method enables effective feature learning without emotion labels. Its main advantages include:

- Zero labelling cost
- No strict alignment between modalities
- Simplified fusion applied only at inference

Despite being unsupervised, this method achieved results that outperformed other unsupervised approaches and even surpassed some supervised state-of-the-art models.

Stage 4: The Causal Evolution (Tracking Sentiment)

AI in affective computing is now moving beyond identifying what emotion is present to understanding why it occurs. To respond empathetically, an AI must reason about emotional causes and track how emotions evolve over time forming a causal chain within a conversation.

4.1 Emotion Cause Detection for Empathetic Response

Gao et al. [5] proposed a framework to improve empathetic response generation by incorporating an emotion reasoner component proposed a model incorporating an emotion reasoner component that enhances empathetic response generation. The model performs two tasks through multi-task learning:

I. Context Emotion Prediction predicts the general emotional tone of a conversation (e.g., "lonely").
II. Emotion Cause Detection identifies the specific words or phrases that triggered the emotion using sequence labelling.

For example, in "I miss my friends; they all live

abroad," the phrase "they all live abroad" is labeled as the cause. The model then passes the cause information to the dialogue generator, making sure that the system focuses more on emotionally important words when generating replies.This was one of the very first attempts to model emotion cause in generation of dialogue.

4.2 Emotional Causal Reasoning in Long-Sequence Multimodal Dialogue

The Cause Motion framework, proposed by Zhang et al. [6], is designed to solve the context problem of model forgetting in long-form conversations. While emotion reasoners identify the cause of the trigger of a particular emotion, they struggle with long conversations where emotional triggers may have occurred many turns earlier. To address this, Zhang et al. introduced Cause Motion, a long-sequence emotional causal reasoning framework.

It combines two main technologies:

- Retrieval-Augmented Generation (RAG): Builds a "dialogue knowledge base" using a sliding window to store past context. When new inputs are added, the system retrieves relevant information from earlier dialogues and context.
- Multimodal Fusion: Integrates audio features like intensity, and speech rate into the knowledge base, allowing RAG to retrieve text as well as audio context for more accuracy.

Integrating RAG with multimodal fusion led to a significant performance gain a GLM-4 model with CauseMotion improved causal accuracy by 8.7% compared to the base model and even outperformed GPT-4.0 by 1.2%.

Evolution ary Stage	Model / Approach	Mod alitie s	Key Objectiv e	Dataset s	Key Metrics
Stage 1: Text Classifica tion	CNN+LS TM (Machov á et al.)	Text	Emotion Classific ation	Kaggle Text Data	F1-score (0.95 for sadness)
Stage 2: Conversa tional (ERC)	BERT-ER C (Qin et al.)	Text	Context-aware ERC	IEMO CAP, MELD	Accuracy (71.70%, 67.11%)
Stage 3: Multimod al (MER)	CM-ROB ERTA (Luo et al.)	Text+ Spee ch	Supervis ed Fusion	MELD	Weighted F1-score (66.28%)
Stage 3: Multimod al (MER)	Unsuperv ised Contrasti ve (Francesc hini etal.)	Text + Audi o + Visio n	Unsuper vised Learnin g	CMU-MOSE I	F1-measu re (69.50%)
Stage 4: Causal (Empathy)	Emotion Reasoner (Gao et al.)	Text	Cause Detectio n	Empat hetic-D ialogue s	Accuracy (42.4%), BLEU, Empathy Rating (3.244)
Stage 4: Causal (Long-Fo rm)	CauseMo tion-GL M-4 (Zhang et al.)	Text +Audi o	Causal Reasoni ng	ATLA S, DiaAS Q	Causal Accuracy (+8.7% over baseline)

Based on the gaps identified in the literature, a project methodology is proposed to build a practical, real-time, text-based conversational agent. This methodology focuses on mastering the foundational and conversational stages (Stage 1 and 2) of the evolution analysed. The project is structured into several key phases

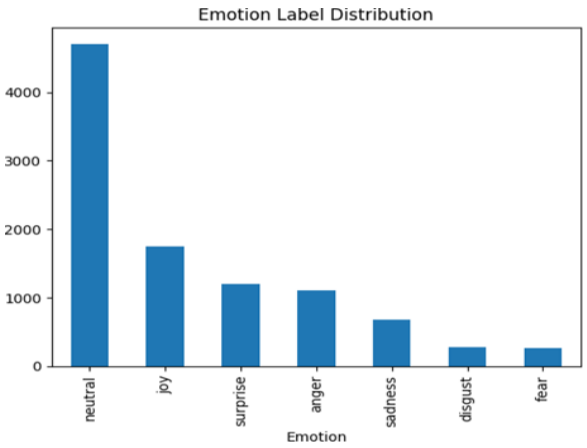
Inference Pipeline		
3.6 Evaluation & Iteration	Measure F1-score, user satisfaction.	Statistical analysis; human-subject studies

III. METHODOLOGY

3.1. Preprocessing

Each utterance from MELD [12] is lowercased, stripped of punctuation, and non-alphabetic characters are removed. Texts are tokenized using each model's default Hugging Face tokenizer: bert-base-uncased, distilbert-base-uncased, and roberta-base. (Note: We acknowledge a limitation in this approach; the roberta-base model is case-sensitive, and lowercasing creates a pre-training/fine-tuning mismatch, which likely impacted its performance on minority classes.) Class labels are index-encoded. Utterances longer than 128

tokens are truncated; shorter ones are padded. Training, validation, and test splits follow the official MELD partitioning.



3.2. Model Architectures

BERT: BERT [13] is a bidirectional transformer encoder. We use the bert-base-uncased model and add a single dense classification head atop the token representation.

DistilBERT: DistilBERT [13] is a lightweight distilled version of BERT. We use distilbert-base-uncased, which offers faster training and inference with a small trade-off in performance.

RoBERTa: RoBERTa [13] modifies BERT's pretraining. We use the roberta-base model, which, as noted, is case-sensitive.

3.3. Fine-Tuning Procedure

All models are initialized from Hugging Face's pretrained checkpoints and fine-tuned for 3 epochs using the AdamW optimizer, learning rate 2×10^{-5}, and a batch size of 8. We use early stopping based on validation macro F1-score, as this is the more relevant metric for our imbalanced, multi-class problem. Training is performed in Google Colab using a T4 GPU.

3.4. Evaluation

Model performance is assessed using Macro F1 (the unweighted mean of F1-scores across all classes) and Accuracy (the percentage of correct predictions). We also report per-class F1-scores.

BERT Results:

BERT shows overall performance with accuracy of 60% and Macro F1 score of 0.483 which is the highest among all the models trained.

Emotion	Precision	Recall	F1-Score	Support
Anger	0.517	0.392	0.446	153
Disgust	0.467	0.318	0.378	22
Fear	0.409	0.225	0.290	40
Joy	0.540	0.534	0.537	163
Neutral	0.704	0.800	0.749	470
Sadness	0.455	0.360	0.402	111
Surprise	0.538	0.620	0.576	150
Accuracy			0.606	1109
Macro Avg	0.519	0.464	0.483	1109
Weighted Avg	0.591	0.606	0.594	1109

DistilBERT Results:

DistilBERT achieves an accuracy of 59% with macro F1 score 0.429, which is lower than BERT. However, it took the lowest time for training. Training time in min:sec of BERT: 13:59, DistilBert:07:13, RoBERTa: 14:42. DistilBERT took almost half the time to train than other models and still gave results which could tackle the other two models, therefore for large scale, a low-cost solution, DistilBERT is a decent option for training.

Emotion	Precision	Recall	F1-Score	Support
Anger	0.445	0.373	0.406	153
Disgust	0.267	0.182	0.216	22
Fear	0.308	0.100	0.151	40
Joy	0.512	0.534	0.523	163
Neutral	0.718	0.811	0.761	470
Sadness	0.434	0.297	0.353	111
Surprise	0.551	0.647	0.595	150
Accuracy			0.598	1109
Macro Avg	0.462	0.420	0.429	1109
Weighted Avg	0.575	0.598	0.581	1109

RoBERTa Results:

RoBERTa attains highest accuracy of 60.7% but the macro F1 score is lowest (0.412) which shows it performed well on majority classes but it faced problems with minority emotions like fear, disgust etc.

Emotion	Precision	Recall	F1-Score	Support
Anger	0.452	0.366	0.404	153
Disgust	0.154	0.091	0.114	22
Fear	0.182	0.050	0.078	40
Joy	0.548	0.528	0.538	163
Neutral	0.702	0.838	0.764	470
Sadness	0.452	0.297	0.359	111

Surprise	0.588	0.667	0.625	150
Accuracy			0.607	1109
Macro Avg	0.440	0.405	0.412	1109
Weighted Avg	0.575	0.607	0.584	1109

IV. RESULTS AND DISCUSSION

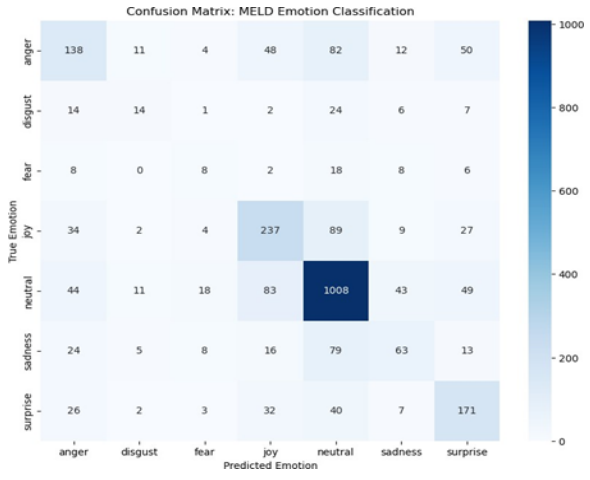
The primary objective of the proposed methodology (Section 3) is to develop a conversational agent capable of recognizing primary emotions (e.g., joy, sadness, anger, fear, surprise, disgust and neutral) from text input. [1].

The evaluation of this system will be conducted using both quantitative metrics and qualitative user feedback, as described in the evaluation phase (Section 3.6). The key Evaluation metrics include:

- F1-Score: It is the measurement of classification performed by the model accuracy by balancing precision and recall into a single value for each emotion class [1].
- User Satisfaction: Evaluated through multiple human-subject studies for perceived empathy, contextual relevance, and overall effectiveness of the agent's emotion-aware responses [1].

The discussion of the result will primarily focus on accuracy, F1-score. So that the final conversational agent is reliable, accurate and responsive enough for real-Conclusion

Confusion Matrix



The journey began with simple text classification, whereby classic machine learning models (Naïve Bayes, SVM) were gradually superseded by high-performance deep learning neural networks (CNN, LSTM) for classifying emotions in discrete text inputs [1].

After that, the ERC, with context awareness, overcame the limitations of its predecessor by incorporating conversational context. This stage made a shift from the traditional two-step model training methods to a more modern, innovative fine-tuning methods of pre-trained language models (BERT-ERC), which models speaker roles, dialogue history, and temporal structure within the fine-tuning process itself [2].

Then to remove the problem of text-only analysis brought about the next stage, leading to Multimodal Emotion Recognition or MER. This phase confronts the "heterogeneity gap" by merging text with audio and video as well. It has been advanced through both supervised frameworks like Cross-Modal RoBERTa [3] and unsupervised frameworks which did not even needed labelled dataset, such as Modality-Pairwise Unsupervised Contrastive Loss, which reduce the need for costly labelled data [4].

The current represents the evolution where the model not only identifies the emotion but also focuses on what was the reason that triggered the emotion. This includes emotion reasoner models, which detect specific the emotional causes and eventually generate empathetic responses [5], and long-sequence causal reasoning frameworks such as CauseMotion, which uses Retrieval-Augmented Generation (RAG) and multimodal integration tracking complex emotional chains over time [6].

The "Emotion and Sentiment Evolution" thus presents a clear trajectory a shift from isolated emotion classification to a causal-aware understanding of human affect. This progression provides the basis for more realistic, efficient, natural, and Empathetic Human–Machine Interaction.

V. CONCLUSION

This paper has analyzed the technological evolution of AI in emotion and sentiment analysis, and its journey with different stages and how it's evolution. We compared BERT, DistBERT and RoBERTa for text-based emotion recognition on the MELD dataset. While RoBERT had the highest accuracy, BERT achieved the best macro F1-score. Hence, we use BERT for chatbot Integration. Among these,

DistBERT provided the best speed accuracy trade-off, therefore for large data DistilBERT is a good enough option. Finally, we show how these classifiers can be integrated into an emotion aware prompting strategy that aims at improving the empathetic quality of LLM chatbots.

VI. FUTURE SCOPE

The analysis of this evolutionary trajectory reveals several key bottlenecks:

(1) the high cost and difficulty of data annotation, especially regarding multimodal and casual data [3], [4];

(2) the "forgetting problem" of standard LLMs in long-sequence dialogues [6]; and

(3) the lack of causal understanding in most classification models [5].

The future scope of this field lies in developing a unified framework that combines the most advanced solutions to these problems. A next-generation model would integrate the key breakthroughs identified in this paper:

Unsupervised Pre-training: The framework would first leverage the unsupervised, modality-pairwise contrastive loss method [4] to pre-train powerful and aligned feature extractors for text, audio, and video, eliminating the dependency on expensive, manually labeled emotion datasets.

Multimodal RAG: These unsupervised, multimodal embeddings would then be used to populate the dialogue knowledge base of a "CauseMotion"-style Retrieval-Augmented Generation (RAG) framework [6].

Causal Reasoning: This would create a system capable of performing long-sequence, multimodal causal reasoning. It could track how a user's tone of voice [3], facial expressions [4], and text from potentially hundreds of turns in the past influence their current emotional state, all without relying on manually labeled causal data [6].

While a text-only project provides a critical foundation by mastering the techniques from the first two evolutionary stages [1], [2], a significant future direction for such work is the extension into these advanced multimodal frameworks [3], [4] and causal frameworks [5], [6].

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Vishwakarma Institute of Technology for providing all the resources and support necessary for this project. We would like to especially thankful to our guide, Prof. Sanyukta Deshmukh, who has guided us through the development of this research with her valuable feedback, comments and words of encouragement throughout the research. [1, 1]

REFERENCES

[1] K. Machová, M. Szabóova, J. Paralič, and J. Mičko, "Detection of emotion by text analysis using machine learning," *Frontiers in Psychology*, vol. 14, p. 1190326, 2023.

[2] X. Qin, Z. Wu, J. Cui, T. Zhang, Y. Li, J. Luan, B. Wang, and L. Wang, "BERT-ERC: Fine-tuning BERT is enough for emotion recognition in conversation," *arXiv preprint arXiv:2301.06745*, 2023.

[3] J. Luo, H. Phan, and J. Reiss, "Cross-modal fusion techniques for utterance-level emotion recognition from text and speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[4] R. Franceschini, E. Fini, C. Beyan, A. Conti, F. Arrigoni, and E. Ricci, "Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss," 2022.

[5] J. Gao, Y. Liu, H. Deng, W. Wang, Y. Cao, J. Du, and R. Xu, "Improving empathetic response generation by recognizing emotion cause in conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 807–819.

[6] Y. Zhang, Y. Li, Z. Yu, F. Tang, Z. Lu, C. Li, K. Dang, and J. Su, "Decoding the flow: CauseMotion for emotional causality analysis in long-form conversations," *arXiv preprint*, 2025.

[7] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogRNN: An attentive RNN for emotion detection in conversations," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6818–6825.

[8] D. Ghosal, N. Majumder, S. Poria, N. Chhaya,

and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.

[9] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, R. Ghosh, N. Chhaya, A. Gelbukh, and R. Mihalcea, "Recognizing emotion cause in conversations," *arXiv preprint arXiv:2012.11820*, 2020.

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[11] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[12] S. Poria *et al*., "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. ACL*, 2019.

[13] L. Romero Gomez, T. Watt, and K. O. Babaagba, "Emotion recognition on social media using natural language processing (NLP) techniques."