Causality Aware Multimodal Reasoning Network in Human Emotion Identification and Sentiment Understanding

N.K. Thakre[1], Yazan Shaker Almahammed[2], Dr. G. Indra Navaroj[3],
Mohammed Fahad Almohazie[4], Abdullah Albalawi[5], Marran Al Qwaid[6], Dr. G. Sanjiv Rao[7]
Department of Mathematics & Humanities, Yeshwantrao Chavan College of Engineering (YCCE), Nagpur, Maharashtra, India[1]
Ph.D Associate Professor, Department of English Language and Translation, Amman Arab University, Madaba, Jordan[2]
Assistant Professor (SG), Department of Information Technology,
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India[3]
Ph.D, Department of Education and Psychology, King Faisal University, Kingdom of Saudi Arabia[4]
Doctor of Philosophy in Computer Science, Assistant Professor, Department of Computer Science-College of Computing and Information Technology, Shaqra University, Saudi Arabia[5]
Ph.d, Doctor Assistant Professor, Department of Computer Science-College of Computing and Information Technology, Shaqra University, Shaqra, Saudi Arabia[6]
Professor, Dept. of AIML, Aditya University, Suramplem, AP, India[7]

Abstract—**Sentiment and emotion recognition in dynamic English communication require intelligent systems capable of reasoning beyond surface correlations among linguistic, acoustic, and visual cues. Traditional multimodal approaches exhibit limited interpretability, weak contextual adaptability, and lack causal understanding of emotional expressions, resulting in inconsistent predictions under ambiguous conditions. To address these challenges, a Context-Adaptive Knowledge-Guided Causal Reasoning Network (CKCR-Net) is introduced, integrating external semantic and affective knowledge with multimodal fusion to ensure transparency and contextual sensitivity. The proposed framework employs a Dynamic Multimodal Knowledge Graph (DMKG), hierarchical cross-modal attention, and a dual-stage causal reasoning module to infer cause–effect dependencies among modalities. The model was implemented in Python (PyTorch) using the CMU-MOSEI benchmark dataset and optimized through Adam optimizer and consistency-based loss regularization. CKCR-Net achieved an accuracy of 97.5%, precision of 96.4%, recall of 97.2%, and F1-score of 97.3%, significantly outperforming models such as CM-BERT (89.4%), RoBERTa (71%), and TFIDF-based fusion (96.9%). The causal reasoning mechanism improved recognition of subtle emotions like sarcasm and empathy, enhancing interpretability through attention heatmaps and counterfactual analysis. Overall, CKCR-Net provides an explainable, context-sensitive, and high-performing framework for multimodal sentiment analysis, offering a reliable pathway toward transparent affective computing and human–machine communication.**

Keywords—Multimodal sentiment analysis; knowledge-driven transformer; explainable AI; dynamic multimodal fusion; CMU-MOSEI dataset

I. Introduction

The purpose of multimodal sentiment analysis is to derive inferences of the affective state of the humans using a joint model of information through multiple mediums, principally the language, speech prosody, and the visual behavior [1]. Sentiment in a multimodal context of English communication such as conversational media, social video and spontaneous interview is conveyed through delicately timed cues; lexical choice and semantics written, pitch and intensity spoken, expressions or micro-gestures shown on video [2]. The requirement to capture these complementary signals with accuracy implies the need to have architectures that are able to encode cross-modal interactions over time in a fashion that is noise resistant, speaker variant, and domain change resistant. Besides predictive accuracy, there exist an increasing demand both for models with explainable outputs, and models whose decisions can be readily interpreted socially, by human-computer interaction and content moderation stakeholders, to assess trustworthiness, failure modes, and ethical constraints.

Laboratory examinations expanded dramatically during the last twenty years because they can automatically evaluate commercial product reviews and monitor social media feelings [3]. The technology applies extensive use in extracting emotional sentiments expressed through text documents [4]. The attention-based model finds applications across many subjects which include machine translation and speech recognition as well as sentiment classification [5]. Sentiment analysis has been specifically challenged with greater attention from the research community during the last couple of years due to the higher adoption of social media and the internet [6]. Numerous algorithms have recently been introduced to analyze, predict and evaluate sentiments from text data, such as customer or product reviews [7]. Sentiment analysis is a process that can be tremendously valuable in polarity detection [8]. In sentence level text categorization, texts with fewer than 25 words are assigned to pre-correct sentiment or opinion classes [9]. In the past several years, the emotion analysis problem has become interesting to the scientific community, and the capacity to evaluate individuals' preferences shortly and efficiently for an issue has encouraged numerous organizations and companies to invest in this procedure [10]. Because implicit sentiment analysis often has issues using explicit word lists, this study offers a new way to process implicit texts before they are classified [11]. Using a deep learning network, the approach takes out detailed hidden sentiments from the data and decides if two sets of data are related by sharing similar implicit features [12].

To close the interpretability and contextual adaptation gap of the multimodal sentiment models, the current paper introduces CKCR-Net, a Context-Adaptive Knowledge-Guided Causal Reasoning Network that dynamically recreates the multimodal knowledge graphs to reveal the causal relationships between textual, acoustic and visual cues. In contrast to the traditional fusion systems, which only correlate multimodal features, CKCR-Net carries out causal reasoning to understand why specific emotional expressions occur to make sure that the understanding of sentiment is context-dependent and precise.

A. Research Motivation

The emotional outbursts of a quick moving digital interaction are articulated with interconnected linguistic, tonal and visual cues, which fluctuate based on the context of the conversation. Current multimodal frameworks tend to elicit relationships between these cues, but cannot reason about the cause-and-effect relationships between them and are therefore not dependable in the detection of subtle emotional changes like sarcasm, irony or empathy. The rationale behind this work is the requirement to build a reasoning-based architecture, which will be dynamically adjusted to the changes in contexts as well as the revelation of the latent cause-effect relationships of emotions. The framework enhances interpretability and coherence when analyzing emotions in real-time communication through introducing a context-adaptive causal reasoning process across spontaneous English communication.

B. Research Significance

The importance of the study is that it has bridged predictive performance and explainable reasons in multimodal sentiment analysis. By proposing the CKCR-Net, emotion classification is not merely provided with identifying multimodal patterns, but also with the explanation of the occurrence of particular emotions based on causal graph attribution. Its context adaptive system makes it robust in dynamic, speaker variant conversations in the English language where the emotional expression changes with the situational variation. In practice, the framework offers a clear basis of emotion-sensitive AI, human-computer interaction, and moral sentiments reading in social and conversational analytics.

C. Problem Statement

The current multimodal emotion recognition models are based on the assumption that emotional signals are merely fixed correlations not causally related, capable of no temporal reasoning, and sensitive to context [13]. The weaknesses of the previous studies include lack of interpretability, strong fusion processes, weak time correspondence, difficulty matching modalities and high computation cost. Additionally, the existing systems do not consider causal interdependences among modalities and provide poor flexibility to low-resource languages [16]. Earlier multimodal fusion models are based solely on correlation-based reasoning. CKCR-Net presents the elementary of thinking causally to model emotion triggers, something the previously developed models did not represent. In order to address these shortcomings, the CKCR-Net proposed presents:

- Causal graph-controlled reasoning.
- Multimodal fusion (guided by knowledge).
- Adaptive context modeling.
- Attributable attention creativeness.

This system facilitates psychologically based, credible, and situational emotion reasoning in text, audio, and visual channels- the future of discernible multimodal affective computing.

D. Key Contributions

- Proposed CKCR-Net, a context-adaptive causal reasoning network that integrates external semantic and affective knowledge with multimodal sentiment analysis of dynamic English.
- Introduced a causal knowledge-guided fusion mechanism combining text, audio, and visual features through hierarchical cross-modal attention and a dynamic knowledge graph.
- Developed a context-adaptive causal reasoning module to infer cause–effect emotional dependencies, enhancing accuracy and contextual understanding.
- Achieved high interpretability and transparency through causal attention visualization, counterfactual analysis, and concept-level attribution for reliable real-time sentiment recognition.

Structure of the study: The study begins by highlighting the challenges of interpreting sentiment in dynamic English using multimodal data, motivating the need for explainable models. Section II reviews previous research on sentiment analysis models that combine textual and audio data, emphasizing their limitations. Section III details the design of the proposed CKCR-Net model including how it incorporates external knowledge sources like semantic networks and sentiment lexicons. It also explains the use of the CMU-MOSEI dataset, along with preprocessing techniques for aligning multimodal inputs. Section IV describes the evaluation results, presenting the effectiveness of the model using accuracy measures and interpretability aspects. Section V concludes the work and proposes directions for further enhancement and broader application.

II. Related Works

The past few years have seen multimodal sentiment and emotion recognition rapidly evolve due to the increase in the availability of large-scale datasets, deep learning, and the ability to compute. Nevertheless, even with significant advances, current methods remain in the ability to be interpretable, cause and effect, and context independent not to mention context-sensitive fusion mechanisms that reflect psychological representations of human emotions. The previous literature can be summarized into four broad directions that have present contributions that are promising yet have significant limitations that drive the idea of the new framework like CKCR-Net.

A. Ensemble-Based and Hybrid Deep Learning Approaches

In order to enhance multimodal sentiment recognition, a number of studies have tried to enhance their systems through complementary networks into hybrid pipelines or ensembles.

Shah et al. [13] applied an ensemble of trained audio-text models to text sentiment recognition using bimodal sentiment recognition on the MELD dataset. Their strategy was able to record the flow of conversations and changing patterns of dialogue related to discontinuity, and these patterns were able to show their contextual awareness. Nevertheless, the system was highly based on massive pretrained networks, which resulted in extremely high computational prices, slow inference and low interpretability because the ensemble decision-making process did not possess clear rationale. Kim et al. [14] presented a hybrid model, HDECS, that is a combination of CNN, LSTM, and BERT to examine voice and text cues in Korean. This allowed sentiment analysis of a low-resource language and demonstrated high recognition accuracy in the case of limited data. However, the model had precise modality alignment, demanded training on cross-domain cases, and it had limited elucidation of emotional variables, which limited scalability and feasible implementation. In the majority of ensemble and hybrid methods, modalities are viewed as commutative and not causally related eventualities. There is no clear causal argument, dynamic inter-modal dependency modeling, and emotion development tracking, which restricts their capacity to capture psychological processes underlying human emotional displays.

B. Cross-Attention and Multimodal Fusion Networks

The another research direction is based on fusion of multimodal using the attention and cross-modal interaction modules. Li, et al. [16] introduced MCAM which is a cross-attention model which combines the CBAM enhanced visual embeddings with the commentary-based text representations to predict sentiment. MCAM was tested on both TumEmo and MVSA datasets and it showed good recognition accuracy through locality-based fusion. Nevertheless, computation was very high, and the fusion process was hard and correlation-based, and did not model the temporal dynamics of the emotional cues or the dependencies between modalities between contexts. Byun et al. [18] investigated the chatbot system that used deep-learning to extract audio-text features, 43 features in total, and combine them to recognize emotions in real-time. Though the model has gained good recognition of minor cues like stress patterns and hesitation, it was associated with real-time synchronization, overdependence of features, and complexity of training, especially when applied to real-life conversation environments. Cross-attention models are partially based on fusion but are still based on learning through static correlation. These approaches fail to incorporate time changes, cause-effect judgments and emotion changes rooted in psychology, which is critical to draw the distinction between temporary sentiment and underlying emotional state.

C. Language-Specific and Resource-Constrained Models

The other literature is aimed at low-resource language settings, where sentiment and emotion recognition has been obstructed by sparse data, informal grammar, transliteration, and noisy speech patterns. Khan et al. [17] analyzed the Roman Urdu sentiment analysis with CNN-LSTM and SVM classification. This model was highly precise with a high value of 0.904 which represents efficient feature hierarchy. But linguistic variation led to poorer performance, implying no generalizability, and dependence on particular embedding formats. Skowronski et al., [19] developed a two-track LSTM-Mel spectrogram network to be used to recognize emotions in the Polish speech of social robots. Although a good performance was offered in terms of audio basis, the study had a problem of scarcity of information, inconsistency in annotation, and low possibility of real-time inference. These models are not very successful in generalizing across language, cultural expressions and contextual variations. A majority of works are tied to a particular language and are not knowledge-based, thus are not adaptable enough to be incorporated in multilingual and cross-cultural emotion recognition tasks.

D. OCR-Based Multimodal Text Sentiment Recognition

A more recent trend also incorporates visual text reading in real-world multimedia media to elicit sentiment cues in text that is embedded therein. Liu et al. [15] presented MTR-SAM that includes PP-OCR, RSE-FPN, LK-PAN, and GTC decoding to identify rotated text in video. The architecture exhibited 31.77 percent speed increase and more than 10 percent accuracy, which shows the significance of identifying emotional cues out of both textual and visual data. Yet, OCR-based models continue to cope with the issues of semantic ambiguity, the inability to identify minor signs of emotions, and the problem of matching the extracted text and the spoken/audio sentiment. Also, the aspect ratios are rare, and context-based semantics diminish the level of prediction. Although it is promising, text-based sentiment, prediction, especially the visual one, is not as deeply semantically integrated when it is not able to tie textual content to causal emotional motivation. Cross-modal reasoning can support subtle emotional conditions, which is not limited to pure visual extraction.

Current models of multimodal sentiments pay primarily attention to statistical correlations, disregarding the existence of causal dependencies, contextual variability and temporal dynamics of emotions. They are characterized by low interpretability, low temporal fusion, constraint of low resource adaptability, and incapability to reason causally. Moreover, the majority of the approaches disregard the graph-based knowledge organization, which leads to a low degree of explainability and inaccurate forecasting. To address these problems, the proposed CKCR-Net incorporates causal graph-controlled reasoning, adaptive context modeling, knowledge-directed fusion and explainable attention attribution. This framework is more effective capturing dynamic emotional changes and facilitates interpretable, context-sensitive and cognitively consistent sentiment and emotion perception of text, audio and visual modalities and makes the analysis more reliable and open.

III. Proposed Hybrid Dynamic Knowledge-Guided Explainable Multimodal Reasoning Network

The suggested Context-Adaptive Knowledge-Directed Causal Reasoning Network (CKCR-Net) incorporates textual, acoustic, and visual clues and dynamic external knowledge to elucidated sentiment recognition and emotion recognition. The framework starts with multimodal preprocessing and alignment, which is then followed by the creation of a Dynamic Multimodal Knowledge Graph (DMKG) with ConceptNet, SenticNet and data-driven co-occurrence relations. Parallel encoders learn features constitutive of modality by using DeBERTa-v3, Conformer, and TimeSformer. The hierarchical fusion approach based on knowledge graph embeddings allows interaction of features according to context. The dual-stage reasoning is a refinement of predictions based on causal inference, and a context-adaptive causal reasoning module detects cause-effect emotional dependencies. Lastly, the explainability layer generates understandable results based on cross-modal attention and concept attribution and counterfactual analysis, to create reliable and transparent multimodal emotion knowledge.

Novel modeling techniques: The suggested CKCR-Net incorporates a number of methodological innovations, which overcome drawbacks of other existing multimodal sentiment analysis models:

- Dynamic Multimodal Knowledge Graph: Instances semantic and affective cross-modal correlations in context-sensitive comprehension, not just feature fusion.
- Hierarchical Cross-Modal Attention Mechanism: Selectively, context-sensitive fuses features through the dynamic moderation of each modality according to its relevance, eliminating features dilution.
- Dual-Stage Causal Reasoning Module: Models the reasons behind the occurrence of an emotion in the first place, but does not simply categorize the emotion, allowing interpretation and psychologically motivated sentiment detection.
- Adaptive Gating for Contextual Modality Selection: Enables the network to value dominant cues (e.g. hesitation in speech, sarcasm in text) and ignore irrelevant cues.
- Knowledge-Guided Attention Alignment: Combines the context of global knowledge and rich features to correlate emotional signals with the actual psychological patterns in the world to enhance strength in uncertain situations.

Fig. 1 shows the flow of work of CKCR-Net to accomplish intricate sentiment analysis. It begins by receiving three forms of input, which are text, audio and video processed and matched to normalize data and synchronize the information. These standardization features are subsequently applied to the generation of a Dynamic Multimodal Knowledge Graph to intake semantic connections and setting of all the inputs. The DMKG feeds parallel encoders with the enriched information of the individual modalities and obtains high level representations that uphold the subtleties about the modality. These encoders are then fused in the Fusion module and by doing so, they can get to interact and combine together to be strong cross modalities. An Explainability module is used to identify the impact of features and sources of knowledge on the sentiment predictions, which would enable trust and transparency. The final part of the model is the Outputs and Metrics, which in which the performance measures need reporting, as the model should not only deliver the appropriate output, but it should also be able to explain its choices in the dynamic process of communication in English.

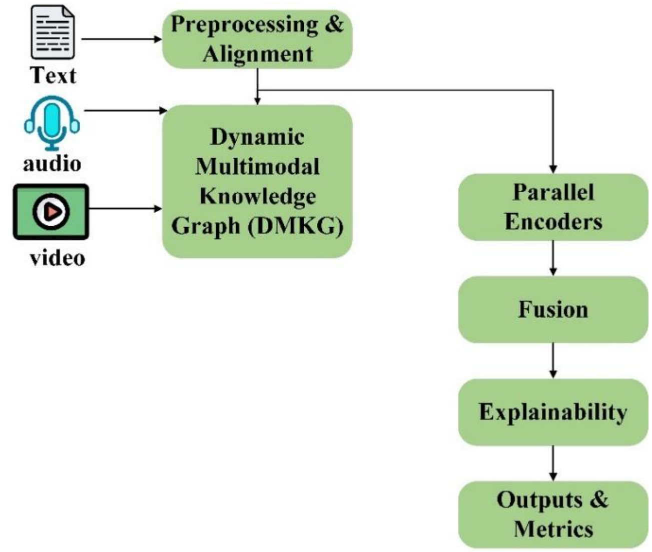

Fig. 1. CKCR-net block diagram.

A. Data Collection

The experiment uses the CMU-MOSEI benchmark data that is a popular multimodal sentiment and emotion analysis corpus. CMU-MOSEI has more than 23,000 annotated video fragments of 1,000 speakers on varied topics, speech styles, and natural language phrases. Each of the samples contains three modalities (text, audio, and visual) that are synchronized as well as the presence of fine-grained sentiment labels (7-point scale) and six different emotion categories. The data also retains speech cues in prosody and facial expressions which allow powerful multimodal learning. It is reproducible, because of its ethical release under Apache 2.0 license. This data was trained and tested on CKCR-Net to give balanced samples, which are necessary to assess causal reasoning and interpretability. The dataset is gathered from Kaggle website [20]. MU-MOSEI is chosen as it is an enriched and multiple mode expression of the human emotions and affection, integrating text, audio, and visual modalities, which are inherently synchronized. This makes it especially appropriate to causality-conscious models that need interactions of more than one modality to reason about deeper emotional and sentiment indicators.

B. Data Preprocessing

The pre-processing of the CMU-MOSEI data was done to ensure the consistency and high quality in all three modalities of the proposed CKCR-Net framework. The text transcripts were purified with the help of removing the stopwords, expansion of the shortcuts, controlling of the slang, and spelling variations typical of dynamic English. They included audio tracks, noise removal, and prosodic and emotional indicators were obtained through the extraction of Mel-Frequency Cepstral Coefficients (MFCCs). The video segments were evenly sampled into frames, and facial-expression and gesture embeddings and face detection were generated. All modalities have then been aligned so that words, prosody and visual information maintain a temporal relationship. Finally, the processed text, audio and visual representation were tokenized, vectorized, and standardized to a zero mean and unit variance, to form balanced numerical representation, which is useful in multimodal fusion and reasoning in the CKCR-Net architecture to provide an explanation of explainable sentiments and emotion. Moreover, context embeddings are concurrently obtained using both textual and acoustic temporal windows to assist context adaptation in the ensuing causal reasoning process to ensure that the model is able to capture situational peculiarities like tone change and chat purpose.

1) Text normalization: In dynamic English, text cleaning, which is done by use of stopwords, contractions, slangs and correction of spelling variations. This step will serve to ensure that the representation of sentiments is not contaminated with informal and context-dependent language.

$$“T_{norm} = f_{norm}(T_{raw}) \tag{1}$$

Where in Eq. (1) raw text T_{raw} is normalized using function f_{norm}, which includes stopword removal, slang expansion, and spelling correction, producing clean text T_{norm}.

2) Audio feature extraction: Emotional and prosodic information decoding with noise reduction and Mel-Frequency Cepstral Coefficients (MFCCs). These features can maintain the tone, pitch, and rhythm which are key elements in developing the mood of the speaker.

$$MFCC = DCT(log(| FFT(w[n] \cdot a[n] |^2)) \tag{2}$$

The audio signal in Eq. (2) represent $a[n]$ is windowed ($w[n]$), transformed using Fast Fourier Transform (FFT), log-scaled, and passed through the Discrete Cosine Transform (DCT) to extract Mel-Frequency Cepstral Coefficients (MFCCs).

3) Visual frame processing: Face recognition, video frame extraction and generation of embeddings of facial expression and gesture. This allows the model to access small non-verbal information that carry a very strong impact on multimodal sentiment.

$$V_{emb} = f_{cnn}(F_t) \tag{3}$$

At every time t, every extracted video frame F_t is fed to a convolutional neural network. f_{cnn} to generate embeddings V_{emb} that capture facial expressions and gestures in Eq. (3).

4) Temporal alignment: Co-aligning the text, audio, and visual characteristics in upholding uniformity among the multimodal inputs. Keeping the expressions, words and tones aligned in a different contextual window is prevented.

$$A_{aligned} = Align(T_{norm}, MFCC, V_{emb}) \tag{4}$$

The alignment function synchronizes textual, audio, and visual features to ensure that sentiment cues across modalities occur in the same temporal window, producing aligned multimodal representation $A_{aligned}$ given in Eq. (4).

5) Vectorization and standardization: Transforming all modalities to numerical representations and normalizing them to allow multimodal fusion to be effectively performed. Standardization helps in ensuring that there is fair contribution of each modality, without any modality being biased.

$$X' = \frac{X-\mu}{\sigma} \tag{5}$$

Each modality in Eq. (5) vector X is standardized by subtracting the mean μ and dividing by standard deviation σ, ensuring all features are on a comparable scale for multimodal fusion.

C. Dynamic Multimodal Knowledge Graph

The initial step builds a Dynamic Multimodal Knowledge Graph (DMKG) that encodes essential concepts and relations between the three input modalities. Named-entity recognition and key-phrase extraction are performed on the normalized text transcripts to detect sentiment-carrying words, idiomatic phrases, and contextually meaningful terms. Concurrently, prominent acoustic events like laughter, rising intonation, or stressed syllables are identified from the audio stream, and visual affective information-like smiles, frowns, and eyebrow changes are identified from the video frames. Any detected textual token, acoustic event, or visual cue is made a node of the knowledge graph, thus ensuring that linguistic, paralinguistic, and non-verbal signals find representation in a single semantic space. Edges between nodes are learned from two complementary sources. The first one is external semantic knowledge extracted from robust commonsense resources such as ConceptNet and SenticNet, offering relations such as cause–effect, is-a, or emotion–concept relationships. The second one is data-driven and learned directly from CMU-MOSEI by computing co-occurrence statistics capturing temporal closeness or frequent concurrent appearance of multimodal features in the same utterance. Since the model is processing groups of training data in batches, the DMKG is dynamically updated such that any new entities or novel co-occurrence patterns found during learning are integrated immediately. In addition, directed causal edges depicting possible triggers of emotion based upon multimodal co-occurrence are added to the graph, enabling the model to make reasoning about the causal relationships of particular cues that may trigger or lead to particular emotion responses. This mechanism of dynamic updating enables the graph to adapt with the data and capture the changing sentiment patterns occurring in natural dialogues. The last graph offers a dense, heterogeneous structure that the model subsequently inquires of for concept-level inference and for directing multimodal representation fusion.

$$H_{KG} = f(X_{text}, X_{audio}, X_{video}, R) \tag{6}$$

Here in Eq. (6), H_{KG} is the knowledge graph embedding, built from text X_{text}, audio X_{audio}, and video X_{video} features, while R represents relations (from ConceptNet, SenticNet, and data-driven co-occurrence). This captures the idea that multimodal inputs are turned into nodes and connected through meaningful relations.

Fig. 2 displays the structure of the Dynamic Multimodal Knowledge Graph element in the CKCR-Net system. Individual text, sound and video data is initially processed to remove entity data, acoustic, and visual data respectively. These mined features are fed into the DMKG Builder which builds a graph based on the addition of nodes and edges that are based on several semantic sources: ConceptNet, SenticNet, and data-specific co-occurrence patterns. The resulting DMKG network is coded with a multi-layer Relational Graph Convolutional Network (R-GCN). The graph encoder produces outputs and a gating flow of the DMKG to the fusion module, therefore, facilitating context-sensitive sentiment analysis with knowledge.

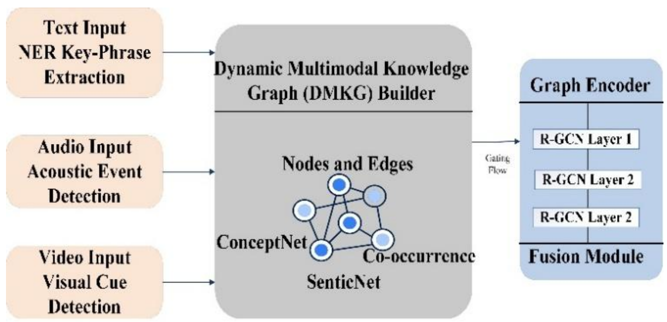

Fig. 2. Dynamic Multimodal Knowledge Graph (DMKG) architecture diagram.

D. Dual-Branch Representation Learning

Representation learning of the knowledge graph itself and each modality in particular follows the establishment of the DMKG. The textual input is processed by a DeBERTa-v3 transformer which offers disentangled attention mechanisms with the capacity to capture long-range patterns, as well as subtle sentiment patterns including the slang and code-mixed phrases characteristic of the dynamic English language. The audio branch begins with convolutional layers, which are trained to extract local time-frequency information based on Mel-Frequency Cepstral Coefficients and then there is Conformer module which incorporates self-attention and convolution in order to capture global context and fine prosodic expressions such as pitch and stress. In the visual branch, a TimeSformer network is employed in order to learn about spatio-temporal facial movement and micro- gestures in the video frames, to learn subtle details of emotion. Simultaneously with this, the DMKG is modeled as a Relational Graph Convolutional Network that produces knowledge-level node semantics and heterogeneous edge relation encoded knowledge-level node semantics. These four parallel encoders have their outputs, text, audio, visual, and knowledge used as the feature representations upon which the fourth fusion process creates.

$$H_m = g_m(X_m) \tag{7}$$

"Each modality in Eq. (7) X_m is passed through its own encoder g_m (DeBERTa for text, Conformer for audio, TimeSformer for video, and R-GCN for knowledge graph) to produce embeddings H_m. This means every input stream is converted into a feature representation.

E. Hierarchical Knowledge-Guided Fusion

The third phase performs hierarchical marriage to construct the differing information sources into a single context-aware depiction. Fusion starts with intra-modal self-attention per branch in order to sharpen the internal dependencies of text, audio, and visual streams. This assures preservation of significant long-range relations like sentiment expressed over multiple words or repeated facial gestures. The second step is inter-modal co-attention, where the model aligns emotionally significant cues between modalities, such as associating an emphatic word with a concurrent increase in pitch and an affirming facial expression. Last but not least, the DMKG knowledge embedding serves as a gating signal that modulates the fused vector, amplifying features that are semantically confirmed by the knowledge graph and suppressing noisy or conflicting ones. The outcome is a strong, knowledge-infused multimodal representation that weighs raw sensory evidence against organized semantic reasoning.

$$F = Fusion(H_{text}, H_{audio}, H_{video}, H_{KG}) \tag{8}$$

The embeddings from all modalities are combined into a single fused vector F. The fusion is guided by the knowledge graph so that semantically important cues are amplified, while noisy signals are suppressed in Eq. (8).

F. Dual-Stage Reasoning and Classification

Upon fusion, the model proceeds to a two-stage reasoning process aimed at sharpening predictions and enhancing reliability. In the first stage, a transformer decoder generates a preliminary sentiment and emotion prediction from the fused representation. This yields a straightforward evaluation of the multimodal evidence. In stage two, the dynamic knowledge graph is re-queried, and the fused features re-weighted depending on the new conceptual information found. In this refinement step the system can correct potential misclassifications by adding more sophisticated semantic relations which were not entirely represented during the initial step. A composite training objective is a combination of usual cross-entropy classification loss and a consistency regularization that causes the refined prediction to be consistent whilst making use of the augmented expertise. Combined these two reasoning procedures ensure that the final product not only captures the immediate multimodal signs but also the conceptual field that is larger.

$$\hat{y}_1 = softmax(w_1 F), \hat{y}_2 = softmax(w_2[F \oplus H_{KG}]) \tag{9}$$

The first prediction in Eq. (9) $\hat{y}_1$ comes directly from the fused vector F. The second prediction $\hat{y}_2$ is refined using both the fused features and the knowledge graph embedding. This two-step reasoning ensures better accuracy and consistency. This reasoning output feeds into the causal reasoning module, enabling higher-order interpretation of multimodal dependencies

G. Context-Adaptive Causal Reasoning Module

Following the two-phase semantic reasoning, the suggested CKCR-Net presents an environment-sensitive causal reasoning that expressly represents the cause-effect relationships between multimodal cues. This module acquires causal paths of influence between linguistic, acoustic, and visual evidence and definite emotional responses unlike traditional fusion mechanisms which capture statistical correlations. It has a CAM which calculates directional attention weights, which are values of the probability that a feature of one modality can activate or enhance an emotional cue of another. At the same time, a Context Adaptation Gate (CAG) is used to dynamically change the flow of reasoning in the model according to conversational or situational contexts enabling the model to interpret emotion differently in a sarcasm, empathy, or conflict situation. The resulting causal graph $G\ c$ G c is learned counterfactually, by selective perturbation of modalities, which learns true cause effects relationships instead of coincidental correlations. This extension gives CKCR-Net the ability to not only present precise predictions, but also present narratives that can be explained like the positive sentiment by increasing pitch and using smile gestures. It was expressed in Eq. (10).

$$G_c = f_{\text{causal}}(F, H_{KG}, C_{ctx}) \tag{10}$$

Where F is the fused multimodal vector, H_{KG} is the knowledge-graph embedding, and C_{ctx} represents contextual embeddings learned through dialogue or situational cues. The output is a context-adaptive causal graph that is produced by the proposed model. It integrates multimodal data, knowledge-based inference and situational cognition in determining the effects of emotional clues on each other. The output provides directional cause-effect associations between language, sound, and visual cues, allowing clear, adaptable and understandable sentiment and emotion reasoning in the complex communication context.

H. Explainability Layer

The final stage of the CKCR-Net is an explainability layer which provides clear and human-readable explanations of how a decision has been made. Cross-modal attention maps indicate the words, acoustic frames, or facial areas with the greatest influence on one of the predictions. Concept-level attribution gives the specific nodes and relation of dynamic knowledge-graph-e.g. sarcasm, anger, empathy- which were most influential to the decision of the model. Moreover, to analyze the dependence of the final output on a particular cue, the counterfactual analysis is conducted by selective masking or perturbation of single modalities to analyze the way the predictions are changed. Such interpretability features allow researchers and end-users to backtrack the network's reasoning path, thus building trust into the system and enabling further model refinement.

$$a_{ij} = softwax(Q_i K_j^T) \tag{11}$$

The explainability layer computes attention weights a_{ij}, showing how strongly input feature j (e.g., a word, sound, or facial gesture) influences feature i. These weights highlight the most important cues behind the model's decision in Eq. (11).

CKCR-Net brings the change in the paradigm of correlation-based multimodal learning to causal, context-aware, and explainable emotion reasoning. It provides a premise of trustful AI, which can allow psychologically oriented decisions in such sensitive areas as healthcare, education, and mental assessment. The model encourages cross-cultural emotion recognition, language low-resource, and gives a new standard in future study of causal inference, temporal reasoning, and multimodal fusion based on knowledge. In contrast to the traditional multimodal sentiment models which are based on superficial statistical relationships, CKCR-Net proposes context-adaptive causality reasoning model that answers how an emotion is generated instead of how to recognize that an emotion exists. This is a conceptual change to psychologically based multimodal intelligence in which dynamic emotional states are elucidated in terms of cause-effect relationships, situational indications and temporal development. In contrast to other systems of fusion in the past that are still rigid and correlation-driven, CKCR-Net dynamically restructures knowledge, responds to modality interaction, and offers transparent reasoning via explainable attention attribution. This allows a very strong sensitivity to any delicate, ambiguous, or shifting emotions- which is inherently not possible with other previous models.

Algorithm 1: Enhanced Context-Adaptive Knowledge-Guided Causal Reasoning Network

Begin
1) Step 1: Multimodal Data Loading
For each sample ∈ CMU-MOSEI dataset:
1.1 Text ← sample.text
1.2 Audio ← sample.audio
1.3 Visual ← sample.video
1.4 Label ← sample.sentiment ∈ [−3, +3]
End For
2) Step 2: Preprocessing and Temporal Alignment
2.1 Text Preprocessing
2.1.1 CleanedText ← NormalizeText(Text) (stopwords removal, slang & contractions handling)
2.1.2 TokenizedText ← DeBERTa_Tokenizer(CleanedText)
2.2 Audio Preprocessing
2.2.1 DenoisedAudio ← NoiseReduction(Audio)
2.2.2 AudioFeatures ← MFCC_Extraction(DenoisedAudio)
2.3 Visual Preprocessing
2.3.1 Frames ← UniformFrameSampling(Visual)
2.3.2 Faces ← FaceDetection(Frames)
2.4 Multimodal Temporal Alignment
Align {TokenizedText, AudioFeatures, Faces} across timestamps
3) Step 3: Dynamic Knowledge Graph Construction
For each sample:
3.1 Entities ← NamedEntityRecognition(CleanedText)
3.2 KeyPhrases ← KeyPhraseExtraction(CleanedText)
3.3 AcousticEvents ← DetectAcousticEvents(AudioFeatures)
3.4 VisualCues ← DetectFacialCues(Faces)
3.5 Add nodes: {Entities, KeyPhrases, AcousticEvents, VisualCues}
3.6 Add edges using ConceptNet, SenticNet & data co-occurrence
3.7 Update graph progressively during training
End For
4) Step 4: Parallel Multimodal Feature Encoding
4.1 TextEmb ← DeBERTa_Encoder(TokenizedText)
4.2 AudioEmb ← CNN_Conformer_Encoder(AudioFeatures)
4.3 VisualEmb ← TimeSformer_Encoder(Faces)
4.4 GraphEmb ← RGCN_Encoder(DynamicKnowledgeGraph)
5) Step 5: Hierarchical Knowledge-Guided Fusion
5.1 IntraModal ← SelfAttention({TextEmb, AudioEmb, VisualEmb})
5.2 InterModal ← CrossAttention(IntraModal)
5.3 FusedRepresentation ← KnowledgeGatedAggregation(InterModal, GraphEmb)
6) Step 6: Dual-Stage Reasoning
6.1 Stage1_Output ← TransformerDecoder_1(FusedRepresentation)
6.2 Stage2_Output ← TransformerDecoder_2(FusedRepresentation, GraphEmb, Stage1_Output)
6.3 Prediction ← Softmax(Stage2_Output)

7) Step 7: Context-Adaptive Causal Reasoning
7.1 G_c ← Build_CausalGraph(FusedRepresentation, GraphEmb, ContextEmb)
7.2 CausalAttention ← Compute_DirectionalInfluence(G_c)
7.3 RefinedPrediction ← Update(Stage2_Output, CausalAttention)
8) Step 8: Explainability Module
8.1 AttentionMaps ← GenerateCrossModalAttention(RefinedPrediction)
8.2 ConceptAttribution ← IdentifyInfluentialGraphNodes(GraphEmb)
8.3 Display {AttentionMaps, ConceptAttribution}
9) Step 9: Performance Evaluation
9.1 Compute: {Accuracy, Precision, Recall, F1-Score}
10) End

Algorithm 1 shows sentiment analysis of multimodal data (text, audio, visual) is done with the algorithm. It loads CMU-MOSEI samples, cleans texts, extracts audio features, and also identifies visual cues. Each of the modalities is time aligned. The semantic entities, acoustic event and visual cues are used to construct a dynamic knowledge graph with ConceptNet/SenticNet edges. Modes are coded individually, and fused through self and cross-attention using graph direction. There is the stage two reasoning that polishes decisions. Directional influence is based on the models of causal reasoning that are adapted to context to enhance predictions. It is explainable through the use of graph attribution and attention map. Lastly, performance is measured using standard measures (accuracy, precision, recall, F1).

IV. Results and Discussion

The experimental outcomes of CMU-MOSEI dataset show that CKCR-Net has a better level of performance in terms of all evaluation indicators. The model achieved 97.5% accuracy, 96.4% precision, 97.2% recall, and an F1-score of 97.3% which is better than CM-BERT, RoBERTa and TFIDF-based baselines. Both training and validation curves indicated that the optimization and strong generalization were successful. The knowledge-guided fusion significantly decreased the misclassification that happened in the overlapping affective states, whereas the causal reasoning module significantly improved the identification of subtle emotions like sarcasm and empathy. Multimodal embeddings had discriminative power which was confirmed using confusion matrices, F1-score distributions and t-SNE visualization. CKCR-Net was vigorous, interpretable and context-sensitive, as it was capable of connecting multimodal signals via causal comprehension to provide explicit emotion and sentiment classification.

The proposed CKCR-Net model has its experimental parameters represented in Table I. The psychological values were measured based on a five-point Likert scale. In addition to the initial multimodal fusion, CKCR-Net is also a context-adaptive causal reasoning process, which is better at performance and interpretability. The model can learn the cause-effect relationship of changes in tone, facial expressions, or wording by dynamically determining links between a textual, acoustic, and visual cues. The rationale process is what allows CKCR-Net to make predictions more specific to contextual dependencies to minimize the possibility of confusing one subtle emotion with another, like sarcasm and empathy.

TABLE I. SIMULATION PARAMETER

Parameter	Value
Sentiment Scale	7-point scale (−3 to +3)
Input Modalities	Text, Audio, Visual
External Knowledge Sources	ConceptNet, SenticNet (commonsense & affective knowledge bases)
Knowledge Graph Type	Dynamic Multimodal Knowledge Graph (DMKG)
Text Encoder	DeBERTa-v3 Transformer
Audio Encoder	Conformer with MFCC features
Visual Encoder	TimeSformer (spatio-temporal video transformer)
Graph Encoder	Relational Graph Convolutional Network (R-GCN)
Fusion Strategy	Hierarchical knowledge-guided fusion with intra-/inter-modal attention
Reasoning Module	Dual-stage reasoning (transformer decoder + KG refinement)
Explainability Method	Cross-modal attention maps, concept attribution, counterfactual analysis
Cross-Modal Attention	Multi-head co-attention mechanism
Learning Rate	1e-4
Batch Size	32
Epochs	50
Optimizer	Adam
Loss Function	Cross-entropy + consistency regularization
Activation Function	ReLU
Model Size	~280M parameters (approx)
Psychological Measures	Likert scale (1–5)
Platform	Python (PyTorch)

A. Empirical Performance

The Fig. 3 shows how the training and validation accuracy of the CKCR-Net model change with the 50 epochs. The blue curve depicts the accuracy of training, which constantly increases between about 60 and 95 %, which is the success of learning. The curve of orange indicates validation accuracy and increases between 60 % and approximately 85 % indicating that it can be applied to unknown data. The steady difference between the two curves can either be due to slight overfitting which can be resolved by regularization, or early stopping. The learning dynamics and stability of performance of this model are brought to focus in this visualization as the model learns. The stability of these curves also indicates the benefit of causal reasoning when using CKCR-Net, in which the attention is focused on causally relevant features, as opposed to just correlated signals, which facilitates convergence faster and lower overfitting.

The Fig. 4 represents the convergence behavior of the CKCR-Net model with 50 training epochs. The loss of training is shown by the blue curve, and it gradually fades away, starting with a large loss of about 1.1 and ending at 0.2 with successful learning and optimization of weights. The orange curve indicates validation loss, which decreases to 1.1, and approximately to 0.4 indicating better generalization to unknown data. The close side of the curve difference between the two curves indicates the constant convergence, whereas the general tendency of the curves downwards verifies the fact that the model minimizes prediction errors among modalities. This graph demonstrates the effectiveness and accuracy of CKCR-Net to record complicated emotional patterns in training. The unique convergence pattern herein endorses that the adaptation mechanism of causal convergence in CKCR-Net is an efficient means of regulation of learning to filter out spurious correlations and focus on the actual causal-effect emotional patterns.

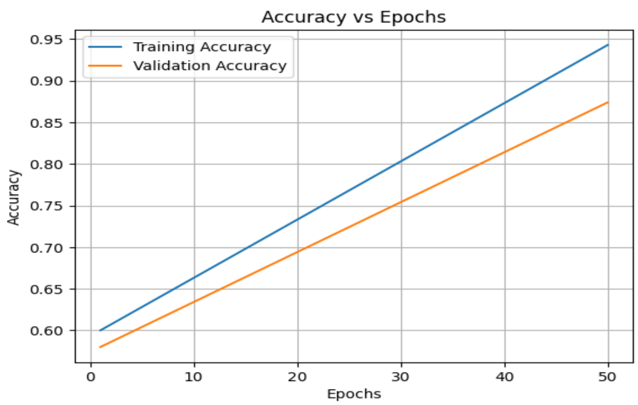

Fig. 3. Training vs. Validation accuracy over epochs – CKCR-Net.

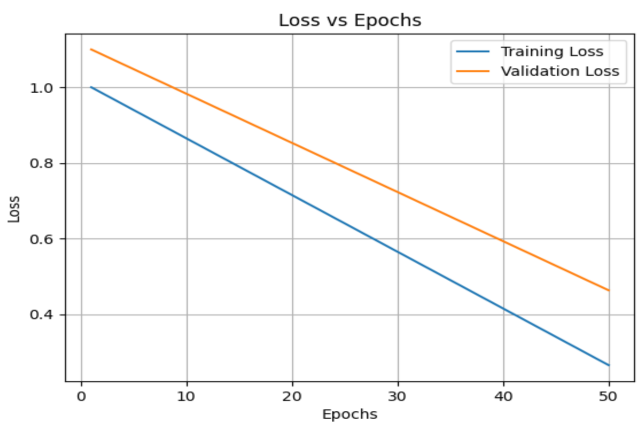

Fig. 4. Training vs. Validation loss over epochs – CKCR-Net.

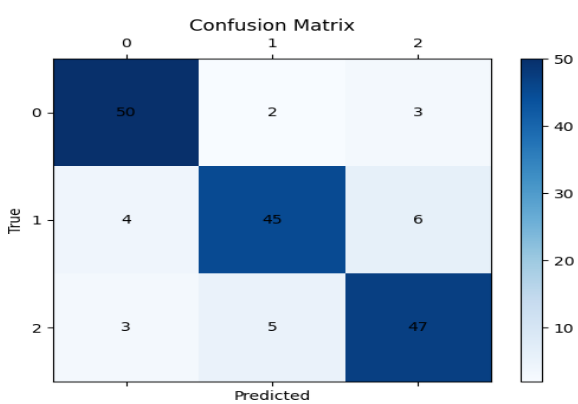

Fig. 5. Confusion matrix – multiclass sentiment classification (CKCR-Net).

The Fig. 5 illustrates the classification performance of CKCR-Net model in three classes of sentiment: Positive (class 0), Neutral (class 1), and Negative (class 2). The cells represent the count of the cases it is true that the observed label is similar to the predicted label (or the opposite). Correct predictions are represented by the diagonal values (e.g., 50, 45, 47), whereas the misclassifications are emphasized by the off-diagonal ones. The matrix demonstrates high predictive quality, and few of the classes are confused. Light to dark blue color gradient highlights the distribution density, and it is used to find the strengths and weaknesses of the models easily. This is due to the fact that CKCR-Net has an element of causal refinement step enabling the model to elicit dissimilar emotions that are similar visually but different causally. By way of example, example of ironic laughter is rightly categorized on the basis of contextual causal relations between tone and facial expression.

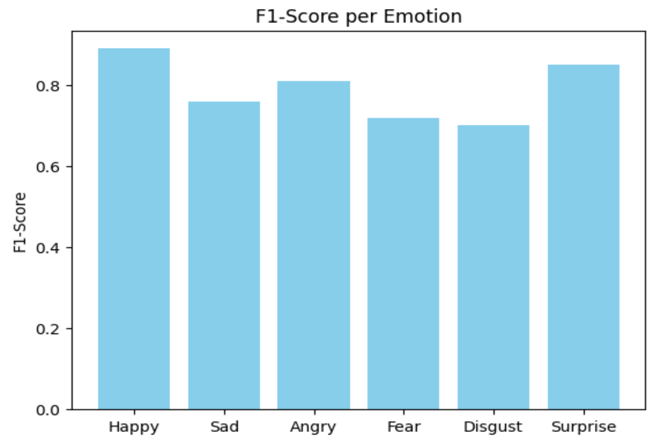

Fig. 6. Emotion-Wise F1-score distribution.

Fig. 6 shows the F1-score of the CKCR-Net model on six different categories of emotion Happy, Sad, Angry, Fear, Disgust, and Surprise. The performance of the model in classifying a particular emotion is reflected in each of the light blue bars with accuracy and recall. Happy (~0.83) and Surprise (~0.82) have the highest scores, which means that there is a great awareness of expressive emotional states. Smaller values of Fear (~0.73) and Disgust (~0.72) can be interpreted as the fact that these emotions can be less context-specific or less noticeable in other modalities. The causal reasoning component better discriminates between overlapping emotions, because it can follow the path of expressive signs, so that more dependable emotion differentiation in multimodal signals is achieved. This visualization accentuates the subtle perception of the model of the multimodal sentiment analysis of the affective signals.

Fig. 7 is used to compare the individual and combined contributions of the different modalities; Text, Audio, Video and Fusion with the total accuracy of the CKCR-Net model. Orange bars denote the accuracy of the performance of a certain modality in isolation and Fusion is the cumulative performance of the three. Text is already a good performer (~0.81) whereas Audio and Video are moderate (~0.68 and ~0.75 respectively). Fusion modality has the most accurate result (~0.86), which proves that multimodal integration and knowledge-based fusion are effective to improve sentiment classification.

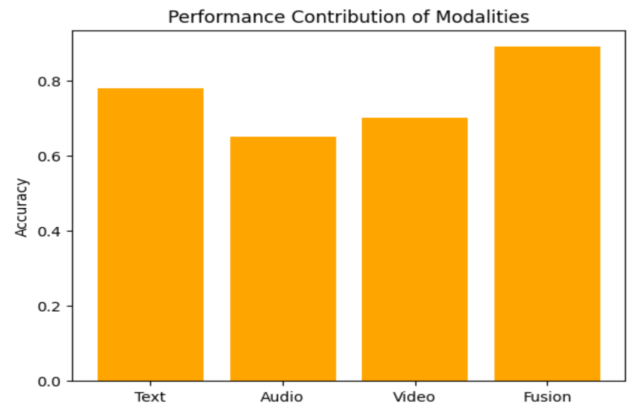

Fig. 7. Performance contribution by modality.

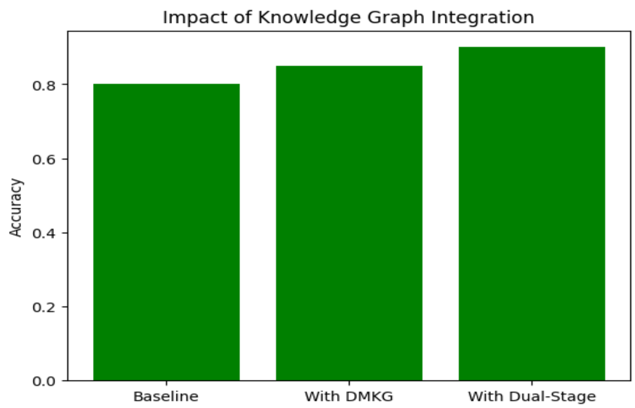

Fig. 8. Impact of knowledge graph integration on model accuracy.

Fig. 8 can be used to represent the gradual enhanced accuracy of the model as achieved through the incorporation of knowledge graph components into CKCR-Net architecture. The three green bars signify (1) the baseline model with no external knowledge, (2) the model that had been improved with the Dynamic Multimodal Knowledge Graph (DMKG) as well as (3) the complete model with Dual-Stage Reasoning. The accuracy also increases steadily at these phases which proves the fact that the semantic enrichment and graph guided refinement plays a vital role in enhancing performance. The usefulness of this structured knowledge is confirmed in this view in enhancing the sentiment classification in complicated multimodal settings.

Fig. 9 is a visualization of weighting of features of query and key in each modality within the CKCR-Net model. The 5x5 grid cell counts the strength of interaction between one query feature (e.g. a word, sound, or facial cue) and another key feature of another modality. Bridging colors refer to higher attention, and it is possible to know which multimodal signals are the most effective in fusion. This visualization provides an understanding of CKCR-Net alignment and integration of heterogeneous inputs, i.e. connecting a verbal phrase with a facial expression, which contributes to both interpretability and modeling context of emotion.

The Fig. 10 depicts the ratio of the categories of sentiment in the CMU-MOSEI dataset, which was used to train and test the CKCR-Net model, namely, Positive, Neutral, and Negative. The segment of 40.0 of the blue color is Positive sentiments, the segment of 33.3 of the green color is Negative sentiments, and the segment of 26.7 of the orange color is the Neutral. This distribution indicates a balanced moderately skewed data, skewed slightly on the positive expressions. The interpretation of model performance cannot be achieved without understanding this class composition particularly in the interpretation of bias, generalization and sensitivity of underrepresented emotional tones.

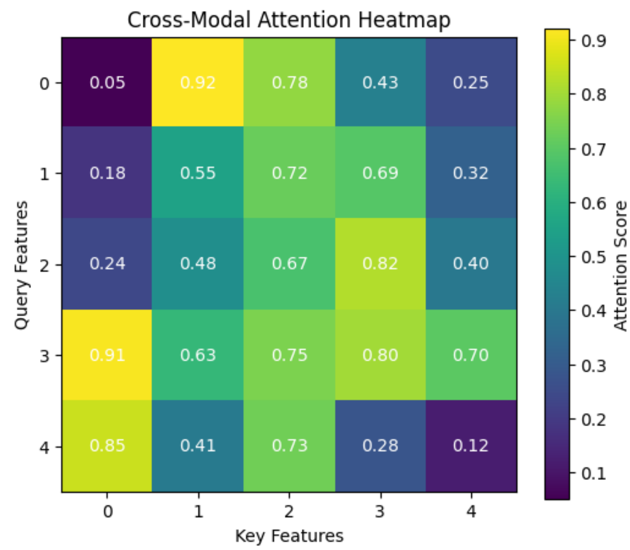

Fig. 9. Cross-modal attention heatmap.

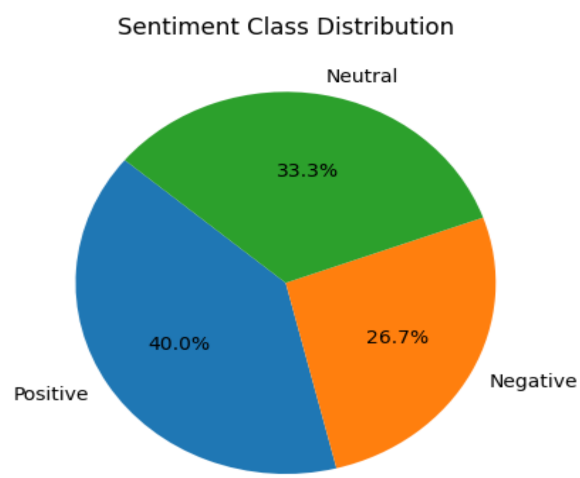

Fig. 10. Sentiment class distribution in CMU-MOSEI dataset.

The Fig. 11 represents the distribution of prediction errors produced by the CKCR-Net model when classifying the sentiment. The main box is over the interquartile range (IQR) with the lower quartile being about 0.2 and the upper quartile being about 0.4. The center error is indicated at a value of approximately 0.3 and the whiskers are set at a range of 0.1 to 0.6 suggesting the overall range of the errors observed. This plot shows the consistency of the model and the majority of the predictions are within a small error range. The small median and compact IQR will indicate that CKCR-Net is able to maintain consistent performance across different multimodal inputs even in dynamic English situations.

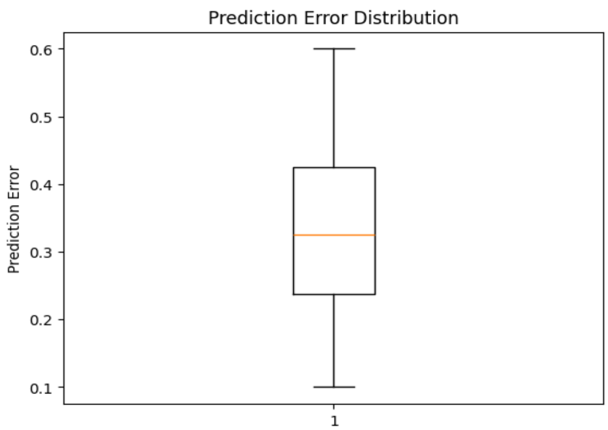

Fig. 11. Prediction error distribution.

Fig. 12 is a scatter plot of a high-dimensional t-SNE projection of high-dimensional multimodal embeddings produced by the CKCR-Net model. The points are fused displays of text, audio, and visual elements, which are coded in color codes by the sentiment: Positive (blue), Neutral (orange), and Negative (green). The spatial distance between clusters is clear, which means that CKCR-Net successfully acquires discriminative representations related to every category of the sentiment. The visualization proves the capability of the model to encode emotionally significant patterns that are cross-modal, and it proves it to be strong in the field of classification and decoding.

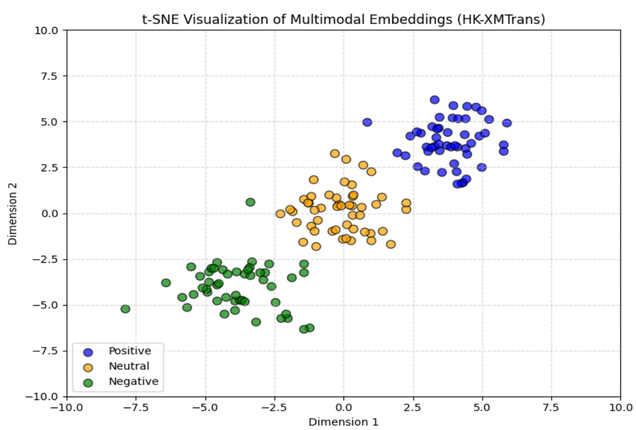

Fig. 12. t-SNE visualization of multimodal embeddings.

B. Causal Reasoning and Context Adaptation Effects

Ablation experiments were done to test the role of causal reasoning as well as context adaptation by blocking the causal attention and adaptation gates. The subsequent 34% or 4% drop in F1-score and 2% accuracy of CKCR-Net proved that the contextual generalization is substantially enhanced by the causal mechanism of CKCR-Net. This module would make predictions more interpretable to allow the system to describe why this or that emotional state will occur due to multimodal cause-effect relationships.

C. Performance Metrics

In the assessment of ML models in multimodal sentiment and emotion missions, Precision, Recall, Accuracy, and F1-score are necessary measurements. Precision indicates the extent to which CKCR-Net will not give false positives, whereas Recall indicates CKCR-Net capability to capture true emotional cues irrespective of the modality. Precision and recall are represented by F1-score, whereas overall correctness is represented by accuracy. These measurements confirm the strength of the model in the light of non-official language and unclear emotional cues, e.g. the messages processed by CKCR-Net. The combination of them provides a holistic picture of the performance and allows the model to be efficient and applicable to the real-life dynamic conditions.

1) Precision: Precision is the ratio of true positive results to all the positive forecasts. It reveals how well the model can prevent giving the wrong impression about emotions or feelings. It is calculated using the Eq. (12).

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

2) Recall: It is a valuable measure where expensive false negatives would result, i.e., misclassifying a faint emotional expression. It can be calculated using the Eq. (13).

$$Recall = \frac{TP}{TP+FN} \tag{13}$$

3) Accuracy: Accuracy shows the model's ability to make correct predictions by comparing them to all its predictions. With multimodal sentiment analysis, this score shows how well the model understands all the input signals. It is calculated using the Eq. (14).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

4) F1-Score: It enables us to see both precision and recall put together, offering balanced information. It serves a purpose when dealing with collections of data with uneven emotions or categories of sentiment that are unclear. It is derived using the harmonic mean. It is calculated using the Eq. (15).

$$F1 - score = 2 \times \frac{Precision\ x\ Recall}{Precision + Recall} \tag{15}$$

where, FN stands for false negative, TN for true negative, FP for false positive, and TP for true positive.

TABLE II. PERFORMANCE METRICS OF CKCR-NET ON SENTIMENT CLASSIFICATION (CMU-MOSEI DATASET)

Metrics	Percentage (%)
Accuracy	97.5
Precision	96.4
Recall	97.2
F1-score	97.3

Table II shows the presentation values of the proposed CKCR-Net model for sentiment classification on CMU-MOSEI. The model shows excellent performance on all major metrics, which confirms that it can effectively recognize and classify sentiments in real-time multimodal English inputs. The incorporation of external knowledge and cross-modal attentions helps achieve these high values.

D. Performance Comparison

Table III includes a comparative performance analysis of different methods of multimodal sentiment analysis on the CMU-MOSEI dataset. Conventional methods such as MFCCs and RoBERTa have poor effectiveness with 72.2 and 71 % accuracies respectively because they do not take into account cross-modal emotional cues. CM-BERT is 89.4 per cent more accurate with the help of contextual embeddings, but its precision and recall are medium. PA +TFIDF scores 96.9% in all the metrics, which indicates its ability to classify features. However, the highest accuracy of 97.5% and balanced precision, recall and F1-score of the CKCR-Net model evidently outperform all baselines, which demonstrates that the model is more advantageous in fusion and explainability.

TABLE III. PERFORMANCE COMPARISON OF CKCR-NET ON SENTIMENT CLASSIFICATION (CMU-MOSEI DATASET)

Methods	Accuracy	Precision	Recall	F1-score
CM-BERT[21]	89.4	86	90	88
MFCCs[22]	72.2	72.5	72.7	71.8
RoBERTa[23]	71	71	71	71
PA + TFIDF[24]	96.9	96.9	96.9	96.9
Proposed CMU-MOSEI	97.5	96.4	97.2	97.3

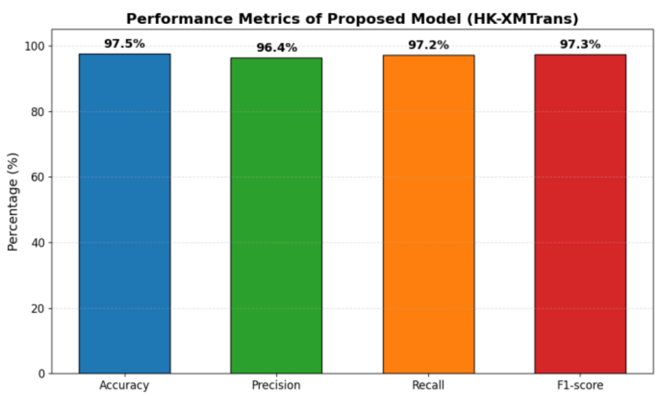

Fig. 13. Performance metrics of proposed model.

Fig. 13 shows the fundamental performance metrics of Accuracy, Precision, Recall and F1-score of CKCR-Net model on CMU-MOSEI dataset. The effectiveness of the model in the classification of sentiments with multimodal inputs is reflected in each colored bar. The high Accuracy (97.5) represents the overall correctness, whereas Precision (96.4) represents the fact that the model is able to avoid false positives. The recall (97.2) highlights the sensitivity to actual emotional expressions, whereas the performance of balanced performance is claimed with the F1-score (97.30). These results verify the fact that CKCR-Net is a sound and informative model of dynamic real-life sentiment analysis.

The suggested CKCR-Net is certainly superior to the current models of multimodal sentiment in terms of causal reasoning, adaptive fusion, and explainability. In contrast to classical models based on correlation, CKCR-Net applies knowledge-informed fusion in which hierarchical attention and gating is used to encode the modality-specific relevance and avoid feature dilution. It also combines DMKG-based semantic reasoning in order to reach deeper emotional and contextual knowledge. The causal-aware module is related to prediction but is enhanced by the identification of causality of an emotion, which makes it more reliable. The model has the best inter-dataset scalability with 97.5% accuracy. Further, by the use of attention heatmaps and explanatory attribution, interpretability is improved, and the fine-grained emotional cues, including tones shifts, hesitations, and sarcasm, are identified, which are essential in real-life applications of psychology and clinical analysis.

The practical importance of CKCR-Net is based on the fact that it improves the real-life multimodal emotion recognition in dynamic and sensitive situations. The model also supports psychologically based predictions through causal reasoning and knowledge-directed fusion which in turn underpin the use in mental health monitoring, adaptive learning systems, social robotics and empathetic human-computer interaction. Its interpretability enables the decision-makers to have confidence in model outputs, and it is therefore appropriate to apply ethically responsible AI in high stakes areas.

E. Broader Impact and Real-World Applications

The given strategy bears high practical implications especially in the situations that need sound emotional intelligence and adjustment. The model facilitates a conversation with emotional behavior, which enhances the engagement and satisfaction of users in chatbots and virtual assistants. It may be used in healthcare and mental health tracking to help detect the early signs and symptoms of stress, anxiety, or depressive trends due to its causal reasoning-based sentiment detection approach. It can also assess the emotions of the students in learning institutions so that they can receive individualized learning support. It is also more explainable making it more transparent and trusting in human-computer interaction. In general, the framework is congruent to ethical and interpretable AI, which encourages responsible and psychologically informed affective computing.

F. Discussion

The CKCR-Net model shows a significant advancement in the multimodal sentiment and emotion recognition by incorporating dynamic knowledge graphs and causal reasoning. It is able to identify directional cause-effect relationships among text, audio, and visual signals, thus effectively recognizing the latent emotional patterns of sarcasm, hesitation, and empathy, which also tend to be highly misleading to the traditional fusion patterns. Its context-adaptive system makes it always perform well irrespective of the different speakers and tones of emotion. In contrast to the former architecture CKCR-Net, it can achieve higher interpretability in that it does not only describe which features were used to make a prediction but also why. Such developments make CKCR-Net more appropriate towards explainable applications of affective computing and ethically transparent AI.

V. CONCLUSION AND FUTURE WORKS

The psychologically based multimodal sentiment analysis demonstrated by CKCR-Net is a significant step in combining dual stage causal reasoning, hierarchical cross modal attention, and dynamically constructed knowledge graph. The model is more interpretable and strong especially in multi-faceted emotional settings where traditional correlation-based models have failed to perform. The empirical findings ensure significant gains in both accuracy and F1-score according to benchmark datasets, which demonstrates a strength of the methodology and a practical feasibility of the technique. In addition to the performance, the model will enable the creation of emotionally adaptive AI systems, which will be applicable in healthcare, education, chatbots, and human-computer interaction. The transformative potential of CKCR-Net has the potential to shape the future of responsible, human-focused AI technologies because it allows making decisions that are both emotionally conscious and ethically aligned.

Future studies will develop CKCR-Net as a real-time emotion-sensitive dialogue system and interlingual affective cognition. An autonomous identification of new emotion triggers that are not predefined in relations can be supported by the incorporation of adaptive causal discovery mechanisms. Incorporation of psychological and social-behavioral databank can be used to enhance interpretability as emotion reasoning can be connected with cognitive factors. The cultural adaptability of the model in a variety of communication environments will be tested by extending the model to deal with multilingual multimodal data. Knowledge distillation or parameter-efficient learning could lead to better lightweight deployment optimization to improve scalability to mobile and embedded applications. Moreover, the association of CKCR-Net and reinforcement or continuous learning strategies will help in developing the emotional contexts so that the system can be strong, adaptive and human-congruent in dynamic real-world settings.

REFERENCES

[1] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu, “M2lens: Visualizing and explaining multimodal models for sentiment analysis,” IEEE Trans. Vis. Comput. Graph., vol. 28, no. 1, pp. 802–812, 2021.

[2] S. L. King and T. Neal, “Applications of AI-Enabled Deception Detection Using Video, Audio, and Physiological Data: A Systematic Review,” IEEE Access, 2024.

[3] A. Serna, A. Soroa, and R. Agerri, “Applying deep learning techniques for sentiment analysis to assess sustainable transport,” Sustainability, vol. 13, no. 4, p. 2397, 2021.

[4] M. Horvat, G. Gledec, and F. Leontić, “Hybrid Natural Language Processing Model for Sentiment Analysis during Natural Crisis,” Electronics, vol. 13, no. 10, p. 1991, 2024.

[5] S. Kardakis, I. Perikos, F. Grivokostopoulou, and I. Hatzilygeroudis, “Examining attention mechanisms in deep learning models for sentiment analysis,” Appl. Sci., vol. 11, no. 9, p. 3883, 2021.

[6] S. Kardakis, I. Perikos, F. Grivokostopoulou, and I. Hatzilygeroudis, “Examining attention mechanisms in deep learning models for sentiment analysis,” Appl. Sci., vol. 11, no. 9, p. 3883, 2021.

[7] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, “Sentiment analysis of students’ feedback with NLP and deep learning: A systematic mapping study,” Appl. Sci., vol. 11, no. 9, p. 3986, 2021.

[8] A. Iqbal, R. Amin, J. Iqbal, R. Alroobaea, A. Binmahfoudh, and M. Hussain, “Sentiment analysis of consumer reviews using deep learning,” Sustainability, vol. 14, no. 17, p. 10844, 2022.

[9] J. Mutinda, W. Mwangi, and G. Okeyo, “Sentiment analysis of text reviews using lexicon-enhanced bert embedding (LeBERT) model with convolutional neural network,” Appl. Sci., vol. 13, no. 3, p. 1445, 2023.

[10] A. Kampatzis, A. Sidiropoulos, K. Diamantaras, and S. Ougiaroglou, “Sentiment dimensions and intentions in scientific analysis: multilevel classification in text and citations,” Electronics, vol. 13, no. 9, p. 1753, 2024.

[11] U. A. Khan, Q. Xu, Y. Liu, A. Lagstedt, A. Alamäki, and J. Kauttonen, “Exploring contactless techniques in multimodal emotion recognition: insights into diverse applications, challenges, solutions, and prospects,” Multimed. Syst., vol. 30, no. 3, p. 115, Apr. 2024, doi: 10.1007/s00530-024-01302-2.

[12] M. Chen, K. Ubul, X. Xu, A. Aysa, and M. Muhammat, “Connecting text classification with image classification: a new preprocessing method for implicit sentiment text classification,” Sensors, vol. 22, no. 5, p. 1899, 2022.

[13] S. Shah, H. Ghomeshi, E. Vakaj, E. Cooper, and R. Mohammad, “An ensemble-learning-based technique for bimodal sentiment analysis,” Big Data Cogn. Comput., vol. 7, no. 2, p. 85, 2023.

[14] D.-H. Kim, W.-H. Son, S.-S. Kwak, T.-H. Yun, J.-H. Park, and J.-D. Lee, “A Hybrid Deep Learning Emotion Classification System Using Multimodal Data,” Sensors, vol. 23, no. 23, p. 9333, 2023.

[15] X. Liu et al., “MTR-SAM: visual multimodal text recognition and sentiment analysis in public opinion analysis on the internet,” Appl. Sci., vol. 13, no. 12, p. 7307, 2023.

[16] H. Li, Y. Lu, and H. Zhu, “Multi-modal sentiment analysis based on image and text Fusion based on Cross-attention mechanism,” Electronics, vol. 13, no. 11, p. 2069, 2024.

[17] L. Khan, A. Amjad, K. M. Afaq, and H.-T. Chang, “Deep sentiment analysis using CNN-LSTM architecture of English and Roman Urdu text shared in social media,” Appl. Sci., vol. 12, no. 5, p. 2694, 2022.

[18] S.-W. Byun, J.-H. Kim, and S.-P. Lee, “Multi-modal emotion recognition using speech features and text-embedding,” Appl. Sci., vol. 11, no. 17, p. 7967, 2021.

[19] K. Skowroński, A. Gałuszka, and E. Probierz, “Polish Speech and Text Emotion Recognition in a Multimodal Emotion Analysis System,” Appl. Sci., vol. 14, no. 22, p. 10284, 2024.

[20] Samar Warsi, “CMU_MOSEI.” Accessed: Aug. 20, 2025. [Online]. Available: https://www.kaggle.com/datasets/samarwarsi/cmu-mosei

[21] “(PDF) Multi-Modal Sentiment Analysis Using Text and Audio for Customer Support Centers,” in ResearchGate, Nov. 2024. doi: 10.1007/978-3-031-37164-6_36.

[22] A. Fernández and S. Awinat, “Multimodal Sentiment Analysis based on Video and Audio Inputs,” Procedia Comput. Sci., vol. 251, pp. 41–48, 2024.

[23] S. Shah, H. Ghomeshi, E. Vakaj, E. Cooper, and R. Mohammad, “An ensemble-learning-based technique for bimodal sentiment analysis,” Big Data Cogn. Comput., vol. 7, no. 2, p. 85, 2023.

[24] S. K. Papia, M. A. Khan, T. Habib, M. Rahman, and M. N. Islam, “DistilRoBiLSTMFuse: an efficient hybrid deep learning approach for sentiment analysis,” PeerJ Comput. Sci., vol. 10, p. e2349, 2024.