applied sciences

MDPI

Article

TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network

Hussein Farooq Tayeb Al-Saadawi and Resul Das *

Technology Faculty, Department of Software Engineering, Firat University Technology, Elazig 23119, Türkiye; husseintaeyb@gmail.com

* Correspondence: rdas@firat.edu.tr or resuldas@gmail.com

Abstract: Affective computing is a multidisciplinary field encompassing artificial intelligence, natural language processing, linguistics, computer science, and social sciences. This field aims to deepen our comprehension and capabilities by deploying inventive algorithms. This article presents a groundbreaking approach, the Cumulative Attribute-Weighted Graph Neural Network, which is innovatively designed to integrate trimodal textual, audio, and visual data from the two multimodal datasets. This method exemplifies its effectiveness in performing comprehensive multimodal sentiment analysis. Our methodology employs vocal inputs to generate speaker embeddings trimodal analysis. Using a weighted graph structure, our model facilitates the efficient integration of these diverse modalities. This approach underscores the interrelated aspects of various emotional indicators. The paper's significant contribution is underscored by its experimental results. Our novel algorithm achieved impressive performance metrics on the CMU-MOSI dataset, with an accuracy of 94% and precision, recall, and F1-scores above 92% for Negative, Neutral, and Positive emotion categories. Similarly, on the IEMOCAP dataset, the algorithm demonstrated its robustness with an overall accuracy of 93%, where exceptionally high precision and recall were noted in the Neutral and Positive categories. These results mark a notable advancement over existing state-of-the-art models, illustrating the potential of our approach in enhancing Sentiment Recognition through the synergistic use of trimodal data. This study's comprehensive analysis and significant results demonstrate the proposed algorithm's effectiveness in nuanced emotional state recognition and pave the way for future advancements in affective computing, emphasizing the value of integrating multimodal data for improved accuracy and robustness.

Keywords: cumulative attribute-weighted graph neural network; trimodal emotion analysis; sentiment analysis; CNN; RNN

Citation: Al-Saadawi, H.F.T.; Das, R. TER-CA-WGNN: Trimodel Emotion Recognition Using Cumulative Attribute-Weighted Graph Neural Network. *Appl. Sci.* **2024**, *14*, 2252. https://doi.org/10.3390/app14062252

Academic Editor: Antonio Fernández-Caballero

Received: 31 December 2023
Revised: 26 February 2024
Accepted: 27 February 2024
Published: 7 March 2024

1. Introduction

Understanding the feelings of other people is an essential skill for day-to-day life. People's responses and behaviors are impacted as soon as their feelings are noticed by other people. If one believes that another person is angry, they should approach them with care [1]. It is essential for a wide variety of sectors and businesses, such as AI, gaming, entertainment, Human-Computer Interaction (HCI), surveillance, and robotics, to have the capability to understand and identify emotions in sensor data [2]. In this assignment, rather than trying to recognize emotions in real life, concentrate on the problem of identifying emotions as they are shown in the assignment. It is vital to collect data from which individuals express, their emotions to construct efficient Artificial Intelligence (AI) systems that can identify emotions [3]. There are many different types of modalities, some examples are body language, gestures, speech, written material, voice modulations, facial expressions, and walking patterns.

Recognizing the influence that human emotions have on our day-to-day lives is essential, which is HCI systems that can recognize these emotions are becoming more

crucial [4]. People operate in a manner that is consistent with their emotional perceptions so that they can respond to environmental stimuli. The ability to comprehend human feelings and behaviors is a capability that is very important for intelligent systems, those used in the fields of medicine, robotics, and surveillance [5]. The first and most important phase in the process of emotion identification is the collection of a wide variety of data formats that represent human emotions. An inherent multi-modality characterizes the manifestation of human emotions to the phenomenon [6]. There are several ways in which a person's emotional state can be communicated, including their voice intonation, pace, facial expressions, word choice, and body language. Based on this, it is reasonable to assume that using a variety of modalities, as opposed to depending on, one can provide more favorable outcomes [7]. The assistant editor who was responsible for reviewing the examination of the article and giving final clearance for publishing was tough to get. Considering that humans are not particularly adept at immediately distinguishing distinct emotional categories in text, audio, or video, it is vital to have evidence that has been accurately labeled [8]. Researchers in emotion recognition have enabled public access to datasets for identifying emotions in both text and images [9]. However, there is a noticeable lack of datasets encompassing all three modalities—audio, text, and video—simultaneously, unlike those that are limited to single or dual modalities. The creation of such comprehensive, trimodal datasets involving video, audio, and text is a complex and costly endeavor [10]. This work aims to evaluate the efficient exploitation of datasets containing varied modal information. The primary objective of this research is to advance the field of Emotion Recognition (ER)by introducing and validating a novel trimodal approach Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN).

Novelty and Contributions of the Study

This study introduces a groundbreaking approach to emotion recognition with the novel Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN) model. Diverging from the conventional focus on primary emotions such as joy, sadness, and fear, our research employs the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSI) datasets. This approach is a significant deviation from existing literature, marking our investigation as the first to apply the CA-WGNN model to these specific datasets, thereby addressing a notable gap in the field.

The innovative edge of our study lies in the CA-WGNN model's ability to capture nuanced emotional states, a facet often overlooked in prior research. By broadening the horizon of emotion recognition beyond traditional categories, we introduce a novel perspective to the domain. The following points encapsulate the core contributions of our study:

- Employing the CA-WGNN model to analyze textual, audio, and video data, using the CMU-MOSI and IEMOCAP databases as trimodal inputs, enhancing the accuracy and scope of emotion recognition (ER).
- Introducing a unique methodology for trimodal emotion identification through the CA-WGNN model, integrating auditory, visual, and textual data.
- Innovating by combining data from multiple modalities through a weighted graph structure within the CA-WGNN model is a significant advancement in multimodal affective computing.

The rest of the paper is organized into sections. Section 2 provides a literature review. Section 3 describes the methods and proposed approach. Performance analysis and discussions are presented in Section 4 and Section 5, respectively. The conclusion of the paper is in Section 6.

2. Literature Review

The study [11] suggested the Deep Learning (DL) technique is used to propose an emotion identification system based on emotional Big Data. Big Data includes audio and video. The suggested system processes voice signals in the frequency domain to create Mel-spectrograms, which can be used as images. Next, the Mel-spectrogram is supplied to a Convolutional Neural Network (CNN). To process video signals, the CNN is fed representative frames from a segment.

The study [12] suggested how well a Long Short-Term Memory (LSTM) mechanism based on DL can identify emotions in text. The research used the Emotion classification dataset, which categorizes emotions into six distinct sets. The findings of the experiment show that, in comparison to other learning approaches, LSTM based text emotion categorization offers considerably greater accuracy. The research [13] proposed a speech-text multimodal emotion detection model to improve emotional recognition system performance. The CNN and LSTM were combined in binary channels to learn acoustic emotion features, while an effective Bidirectional-LSTM network was used to capture textual features. Evolution of the Cumulative Attribute Approach in Computer Vision Regression initiated by [14] in 2013, the Cumulative Attribute (CA) approach in computer vision regression has significantly evolved, addressing complex challenges in fields like age and crowd density estimation. This novel approach is centered around mapping sparse and imbalanced low-level visual features to a 'cumulative attribute space', where each dimension is semantically meaningful and reflects the continuous nature of scalar outputs such as age or people count. By ensuring both discriminative clarity and cumulative conditioning among attributes, CA effectively utilizes the correlation among adjacent scalar values, showcasing substantial improvements in accuracy and efficiency, especially in scenarios with limited and unevenly distributed training data.

The study [15] introduced a new kind of Deep Neural Networks (DNN) that can use text, video, and audio to recognize emotions. Aiming to learn the representation for each modality and the best-combined representation to get the best prediction, the proposed DNN architecture's independent and shared layers work together. Building upon this foundation, further extended the CA approach in 2019 to multifaceted tasks like head pose estimation and color constancy [16]. The essence of CA in this expansion lies in its ability to innovatively map input features to cumulative attributes, capturing intricate correlations between various target values. This method overcomes the limitations of traditional regression models by addressing the interdependencies they often ignore, leading to a more holistic and accurate analysis of multivariate data. Originally effective in single-output problems, the extension of CA into multi-output regression underlines its versatility and potential in broader applications. The implementation of CA involves a two-stage process: the initial stage focuses on attribute learning—mapping from the feature space to the CA space, followed by the second stage, which maps these attributes to the output space. This approach significantly enhances the capability of regression models in visual tasks by leveraging the target space and employing attribute functions corresponding to various target variables.

The article [17] focuses on evaluating various pre-trained language models' performance and environmental impact in text analysis. It highlights the significant trade-offs between model efficiency and ecological footprint in natural language processing, underlining the importance of environmental considerations in developing and applying these technologies in sentiment analysis. The study [18] suggested the methods also neglect merging features from many sources for use in machine learning. Research should focus on developing LSTM-based models capable of handling multi-modal feature fusion and fully accounting for utterance relations. The study [19] proposed a new method for face emotion identification that is based on DL. To start with, the acquired dataset is noise-removed using a combined trilateral filter. After that, the filtered images are enhanced by using contrast-limited adaptive histogram equalization (CLAHE). The study [20] examined to HCI, Speech Emotion Recognition (SER) is a hot subject of study. To assess the performance

of various Artificial Neural Networks models, the study trains one utilizing Mel-Frequency Cepstral Coefficients (MFCC)s feature extraction and tests it on certain audio datasets.

The paper [21] presented an approach that captures the interlocutor and contextual states between utterances using a recurrent neural network (RNN). Before fusing, the modalities' relationships and relative significance are understood via the pairwise attention process. The research [22] examined focuses on the topic of bias in emotion detection systems and how it relates to the modalities used. It examines the effects of multimodal techniques on system fairness and bias. The study [23] suggested that depth and heat detection models are trained using state-of-the-art DL techniques. To analyze publically accessible data in combination for the purpose, this is used in the training process to provide the appropriate annotations for a learning process. The article [24] maintained potential uses in HCI, ERhas lately garnered a lot of attention. Many different spoken and non-verbal languages, such as text, images, and audio, are used to convey human emotions. So, rather than an issue for single-modal learning, emotion detection works well as a multi-modal one. The study [25] investigated image, text, and tag-based emotion identification in this study. Nevertheless, there is an additional, disregarded obstacle that might arise from using many modalities, and that is the missing modality dilemma. Because not all social media posts have text, images, and tags, it's not uncommon for tests to be missing one or two modalities.

The study [26] suggested Attention-based Multimodal Sentiment Analysis and ERto address these difficulties. This paradigm uses visual, auditory, and textual intra- and inter-modality discrimination. Focusing on important task aspects helps categorize sentiment and emotion from visual, spoken, and audio inputs. Modality-specific algorithms automatically extract semantic phrases, image regions, and audio features.

The utilization of Graph operations Graph Neural Networks (GNNs) has emerged as a pivotal element in the field of affective computing, particularly in the realm of emotion recognition. Alsaadawi and Das [27] underscore this significance through their innovative approach in this area. Their research highlights how GNNs adeptly capture the intricate inter-modal relationships within multimodal data, thereby facilitating more accurate and nuanced emotion detection. This advancement represents a significant leap forward in understanding and interpreting complex emotional states using computational methods.

The study [28] examined the Proactive service recommendation driven by multimodal emotion recognition (PSRMER) is a method that the study suggests as a solution to this problem. Using a Trimodal Emotion Recognition (TER) model that is based on Transformer and PSRMER actively recognizes the user's emotion first. The study [29] suggested as more and more researchers focus on the possibility of strong emotional bonds between humans and computers, there is a growing demand for reliable and practical methods for detecting people's emotional reactions. The study presents an EEG recognition model that uses a bandpass filter to pre-process the input signal. The study [30] suggested the process of recognizing and comprehending emotions via the integration of input from several modalities, including text, image, and audio, is known as TER. Nevertheless, a major obstacle to this effort is the lack of labeled data. To do this, this research suggests a unique strategy that combines label correction techniques with consensus decision-making inside a semi-supervised learning framework. The paper [31] suggested highlights the dangers of using DL models with opaque architectures for crucial tasks like emotion identification and explains the fundamental need for human-readable explanations of the model's inner workings.

The article [32] proposed MSMDFN, or multi-stage multimodal dynamical fusion network, aims to address this issue. The MSMDFN serves to acquire the cross-modal correlation-based joint representation. The article [33] discussed emotion categorization using a deep sequential model and a representative characteristic called Complex Mel Frequency Cepstral Coefficients (c-MFCC). To account for small differences in the underlying phonemes, the experimental design does not rely on the speakers involved. The study [34] approached TERcalled Contextualized Graph Neural Networks (COGMEN) is presented

in the research. This system makes use of both local and global information and context. GNNs are the building blocks of the suggested model, which attempts to capture all the interdependencies in a dialogue.

The study [35] examined the recent advances in emotion identification and instances of their use are also included in this overview. Additionally, this poll evaluates several emotion detection sensors and contrasts their pros and cons. Researchers can learn more about current emotion detection systems with the aid of the suggested survey, which will make it easier to choose appropriate sensors, algorithms, and datasets. The study [36] suggested multimodal emotion identification algorithms are receiving more attention from researchers; it is unclear how certain settings might benefit from the combination of visual and non-visual data for emotion detection. The research examines the interaction between two important contextual elements and multimodal emotion components collected from text, tone, and facial expressions. The study [37] suggested that video emotion recognition is plagued by multimodal feature fusion. Fusing mode-specific feature matrices through neural networks is becoming common in DL. Multimodal analysis requires correlations and effective Unimodal features, unlike Unimodal issues. The study [38] examined the need for human-understandable interpretations of black-box DL models are highlighted in this paper, which focuses on the hazards of employing these algorithms for vital tasks like emotion identification.

The article [39] examined low light, non-positive faces, and facial occlusion is a few of the natural challenges that make dynamic expression identification difficult in the field. Approaches that depend on vision can fall short when it comes to capturing the intricacies of human emotions and the framework can efficiently extract multimodal information and provide considerable improvement, allowing to tackle this problem. The study [40] suggested to understanding of emotions grows and new technologies become available, the work builds a reliable emotion identification system that can be used in real-world settings. The article [41] three corpus containing e-commerce data belonging to different languages such as Turkish, Arabic, and English were created and the performances of deep learning and machine learning methods were examined comparatively by performing sentiment analysis on them. The study [42] offered a unique multimodal emotion identification framework that names multimodal emotion identification based on cascaded multichannel and hierarchical fusion (CMC-HF). This framework makes use of visual, voice, and text data as multimodal inputs concurrently.

3. Proposed Approach for Emotion Recognition

Exploring the intricate domain of emotion analysis, our research emphasizes precision and depth, utilizing sophisticated computational methods to unravel and interpret the complex spectrum of human emotions. We implement an extensive, multi-modal approach that merges text, audio, and video analyses for the accurate identification of emotions.

This methodology includes the extraction and processing of textual information via speaker embedding in LSTMs, a key aspect in grasping the nuances of emotional language. Simultaneously, audio features are analyzed using DNN, which discerns patterns in intonation and cadence, essential for understanding emotional states. Complementing this, video features are scrutinized using CNNs to interpret facial expressions and body language, key components in emotion detection. In the audio embedding phase, our approach utilizes MFCC [43], frequency domain linear prediction (FDLP), and wav2vec-based self-supervised speech representations. This integrated strategy enhances speaker-specific information extraction while minimizing extraneous data such as noise. Each feature is processed through independent branches, converging at a shared segment-level layer, thereby reinforcing speaker discrimination. This architecture is particularly adept at improving noise resilience and speaker verification performance in challenging acoustic environments.

The visualizations in Figure 1 provide a dual perspective. Initially, raw audio features are displayed through MFCCs, illustrating the variety in coefficients extracted from audio samples. This is followed by post-DNN embedding visualizations, which demonstrate

the network's capability to reduce dimensionality. The result is a set of 3D embeddings that effectively encapsulate the original MFCCs for accurate speaker verification, thus highlighting the efficiency of our multi-modal approach in emotion analysis.

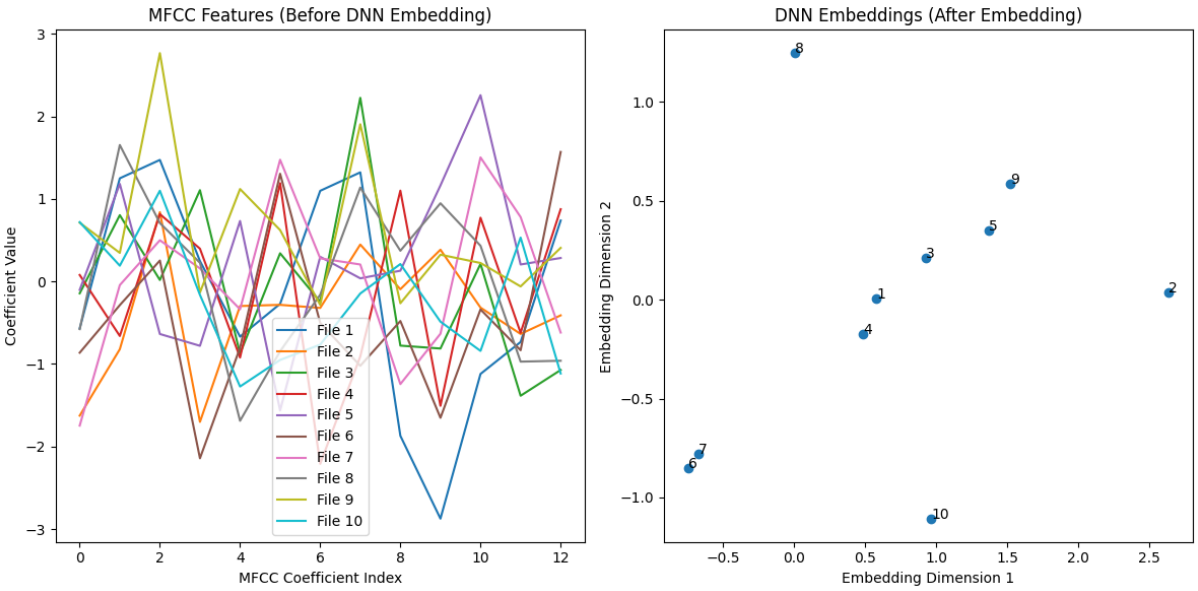

Figure 1. Graphic flow.

As illustrated in Figure 2, our data processing framework utilizes a central dataset that is analyzed through three different neural network models for each type of data. Text data is processed using LSTM, emphasizing the sequential or temporal aspects of text. DNN are used for processing audio data, focusing on extracting complex features such as intonation and cadence, while CNNs are employed for video data, targeting the analysis of facial expressions and body language.

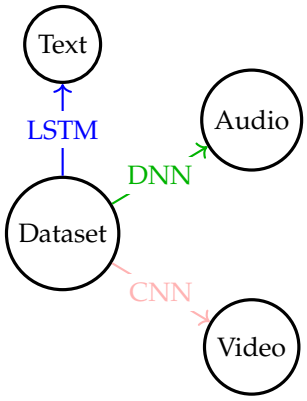

Figure 2. Multimodal data processing framework.

Following the feature extraction phase, we introduce the Cumulative Attribute-Weighted Graph Neural Network CA-WGNN. This forms the core of our trimodal emotion identification system, effectively integrating data from text, audio, and video sources. For a comprehensive understanding of our methodology, please refer to Figure 3. This figure outlines the flow of our process, providing visual insight into how each component synergistically contributes to emotion identification.

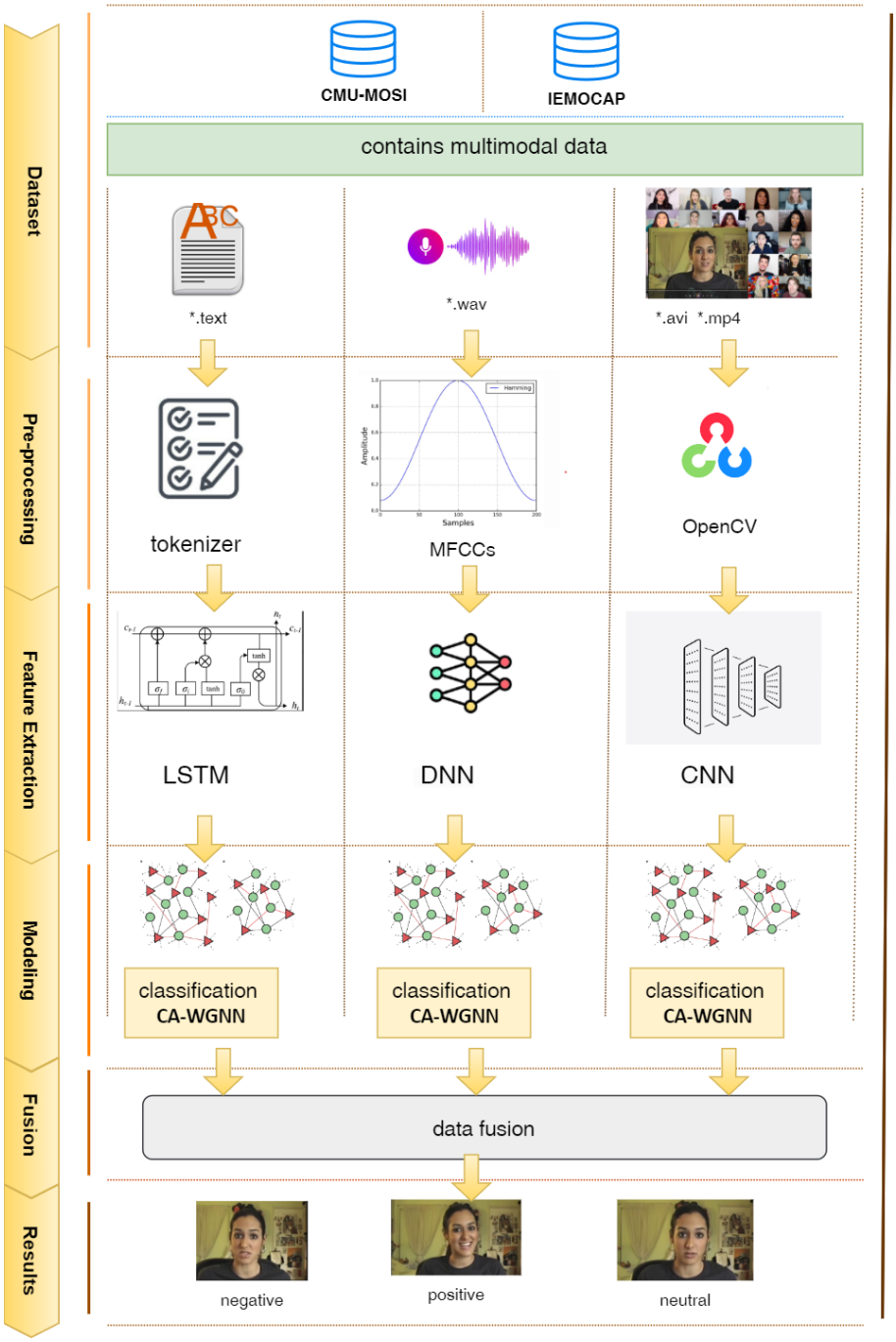

Figure 3. Proposed approach flow chart.

3.1. Dataset Description

Both academics and businesses are increasingly drawn to the study of subjectivity and emotion in online opinion videos. This emerging field aims to understand the nuanced sentiments expressed in these digital narratives. Unlike the more explored domain of text-based sentiment analysis, video, and multimedia sentiment analysis offers a unique challenge. It requires delving into the complex interplay of visual cues, vocal tones, and spoken words to discern underlying emotions and perspectives. The primary hurdle in advancing this field is the lack of a comprehensive dataset, along with established methodologies, baselines, and a thorough statistical analysis of the interactions between different modalities of data. To address this gap, a significant dataset was compiled and is (multicomp.cs.cmu.edu accessed on 26 February 2024) [44]. This dataset serves as a

cornerstone for researchers and practitioners alike, enabling them to explore and analyze the rich tapestry of emotions and subjectivity present in online opinion videos. With this dataset, we can understand the emotional landscape of these videos and develop and refine multimodal sentiment analysis tools.

3.1.1. Dataset IEMOCPA

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [45], originating from the SAIL lab at USC, represents a comprehensive database that is a goldmine for researchers delving into the nuances of human emotions. Spanning twelve hours of rich data, it includes video, audio, facial motion capture, and text transcriptions, offering a holistic view of emotional expression. This diverse dataset transcends traditional boundaries by encompassing text, voice, images, and physical signals, thereby providing a multifaceted perspective on emotional communication. Annotated with both dimensional and categorical labels, the database focuses on dyadic encounters designed to naturally elicit a range of emotions. This approach ensures the capture of genuine emotional responses in a controlled setting, making the data highly valuable for nuanced analysis. For scholars and practitioners in the field of multi-modal expressive human communication, the IEMOCAP dataset stands as an invaluable resource. It not only offers insights into the complex world of emotional expression but also provides a foundational platform for developing advanced Sentiment Recognition systems.

3.1.2. Dataset CMU-MOSI

The Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity Dataset (CMU-MOSI) [44] dataset takes center stage in the realm of Sentiment Recognition, showcasing the power of multimodal sentiment analysis. This innovative dataset is a treasure trove of emotional insights, capturing a diverse range of emotions through spoken words. It does so by skillfully combining visual and auditory data, thus offering a holistic view of human sentiment expression. What sets CMU-MOSI apart is its comprehensive labeling of a wide array of emotions. This feature makes it an invaluable asset for developing and refining emotion identification models, particularly for applications in real-world communication scenarios. The dataset's strength lies in its ability to provide a three-dimensional view of sentiment analysis by integrating text, audio, and video data.

3.2. Speaker Embedding

The journey of analyzing emotions in multimodal data begins with a meticulous feature extraction process, encompassing text, audio, and video modalities. For textual data, the method of speaker embedding in LSTM networks is employed. This technique is adept at capturing the subtle linguistic nuances that are often intertwined with emotional expressions, thereby providing a deep understanding of the spoken word's emotional undertones. When it comes to audio data, DNN come into play. These networks are fine-tuned to extract audio features, specifically targeting the unique patterns found in speech intonation and cadence. This aspect of feature extraction is crucial, as the way we speak often carries a wealth of emotional information, sometimes even more telling than the words themselves. For the video modality, the focus shifts to CNNs. These networks are particularly effective in parsing through visual data to identify key emotional indicators. They analyze facial expressions and body language, which are vital components of non-verbal communication. These visual cues are often the most direct and powerful conveyors of emotion, and their accurate interpretation is essential for a comprehensive understanding of multimodal sentiment analysis.

3.2.1. Text Speaker Embedding Using LSTM

An LSTM network is a marvel in the field of neural networks, distinguishing itself with its unique structure. Unlike traditional neural networks, LSTM are designed with interconnected nodes in their hidden layers. This architecture allows for the sequential

flow of information, where the output of one layer seamlessly feeds into the input of the next. Such a configuration is particularly adept at handling time series data, which is a common characteristic of textual information in speech. The heart of the LSTM lies in its hidden layer, where time series data find their temporal abode. Figure 4 provides a visual representation of the LSTM's fundamental design [46], illustrating how it adeptly manages data over time.

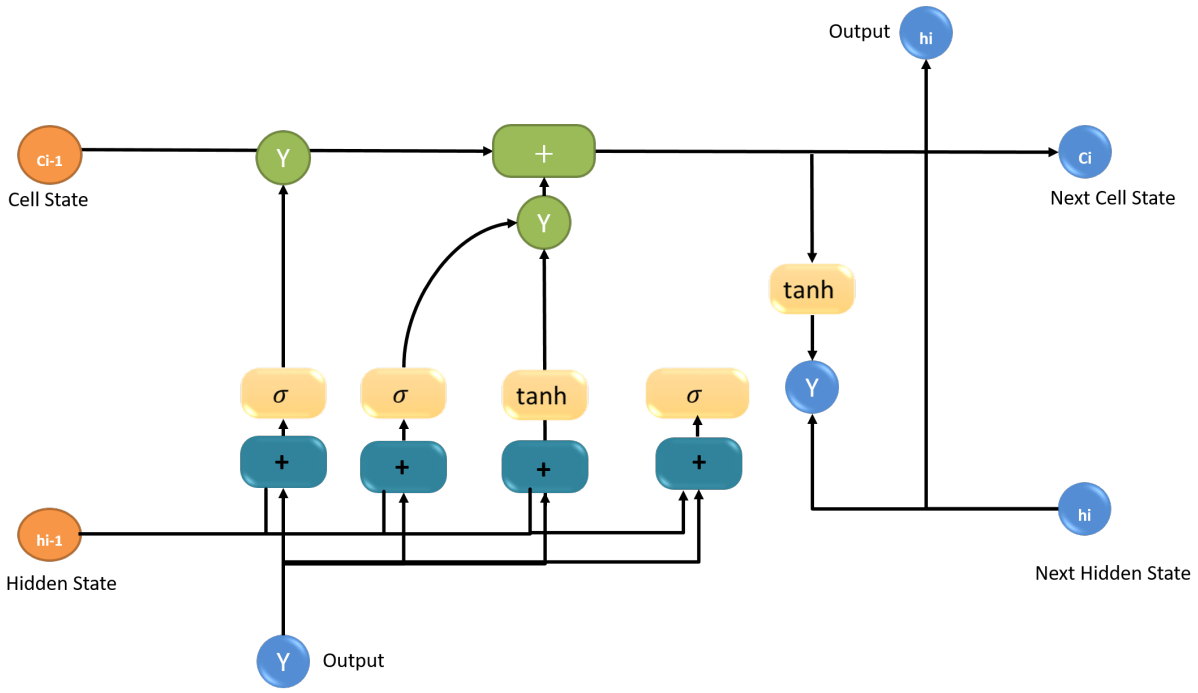

Figure 4. Structure of LSTM method.

A critical component of the LSTM is the 'cell', the basic unit of memory within the network. Each cell is responsible for transmitting not only the cell state but also the hidden state to the subsequent cell, creating a continuous flow of information. This flow of text data through the LSTM chain is unique. The data can traverse through the cell state with minimal or no changes, although some linear transformations may occur. The real magic of LSTM lies in its use of sigmoid gates. These gates act as regulators, updating cell states by selectively adding or deleting text data. The operation of these gates involves a series of matrix operations, each with its independent weight selection, akin to a gate mechanism in traditional computing. One of the most significant advantages of LSTMs is their ability to circumvent the problem of long-term dependencies. Traditional neural networks often struggle with retaining information over extended sequences, but LSTMs overcome this hurdle. They employ gates to manage the network's memory effectively, ensuring that vital information is retained and irrelevant data is discarded, thus maintaining the integrity and relevance of the emotional cues encoded within the text.

3.2.2. The LSTM Network Mechanics

The initial phase in constructing an LSTM network involves identifying the textual data elements that are dispensable. This decision-making process employs the sigmoid function and hinges on the output of the preceding LSTM unit, g_{s-1}, at the $(s-1)th$ position and the current input W_s. The sigmoid function plays a pivotal role in determining which segments of the previous output should be omitted. This gate, termed the forget gate (or u_i), generates a vector with each element ranging between 0 and 1, corresponding to the respective elements in the cell state h_{i-1} Equation (1):

$$u_s = \sigma(Y_u[h_{i-1}, V_i] + b_u) \quad (1)$$

Here , σ denotes the sigmoid function. The matrices Y_u and V_i, along with the bias b_u, are integral to the forget gate's operation. Subsequently, the network decides on retaining and updating the text data within the new input V_i, influencing the modification and evolution of the cell state. This process involves both the sigmoid and tanh layers. The sigmoid layer initially determines the relevancy of new information for updating (0 or 1). The cell state is updated by multiplying these values (see Equation (2)–(4)):

$$q_i = \sigma(Y_q[h_{i-1}, V_i] + b_q) \tag{2}$$

$$N_i = \tanh(Y_n[h_{i-1}, V_i] + b_n) \tag{3}$$

$$C_i = C_{i-1} \cdot u_i + N_i \cdot q_i \tag{4}$$

The variables C_{i-1} and C_i represent the previous and current cell states, respectively. The ultimate output kt is influenced by the cell states' output condition O_i Equation (5), processed through a sigmoid layer that filters the essential components for output. The final cell state C_i undergoes a transformation through the tanh function, ensuring the output values range between -1 and 1 Equation (6).

$$O_i = \sigma(Y_o[h_{i-1}, V_i] + b_o) \tag{5}$$

$$h_i = O_i \cdot \tanh(c_i) \tag{6}$$

3.2.3. Audio Speaker Embedding Using DNN

DNN algorithms have been extensively utilized in both industrial applications and research in recent years, demonstrating a high success rate in recognition tasks. DNN are particularly effective in multi-label classification problems. This study focuses on distinguishing between two distinct groups: the swing and stance phases of gait. The dataset for training encompasses five parameters related to gait. A typical DNN architecture comprises an input layer, an output layer, and at least two hidden layers. Key components of the network include the activation function, learning rate, gradient descent optimizer, and the stages of training. The optimizer refines the process of gradient descent. One notable aspect of DNN is their implementation of the softmax function, as shown in Equation (7), which normalizes neural outputs to the (0, 1) interval, making it highly useful for multi-classification tasks. In these tasks, the model's prediction target is assigned to the output node with the highest probability:

$$r_h = \frac{d^h}{\sum_{u=1}^{b} d^u} \tag{7}$$

The execution time of the method is linearly positively correlated with the encoding dimension, which in turn is influenced by the number of deployment units in a DNN. Following a layer split, the deployment units in a DNN are reduced, leading to a potential increase in execution time and a decrease in the encoding dimension, especially when considering offloading issues.

3.2.4. Video Speaker Embedding Using Deep Learning Architectures

Inspired by neuronal structures in the cerebral cortex, DL architectures have demonstrated efficacy across a variety of machine learning applications. These architectures operate as feed-forward networks, enabling unidirectional information propagation from input to output layers. The design and functionality of these systems draw heavily from biological processes, particularly emulating the hierarchical organization of the visual cortex composed of simple and complex cell layers. Despite variations in specific architectural designs, a common framework exists, comprising a sequence of convolutional and pooling layers, followed by densely connected layers. This design is reminiscent of traditional neural networks. The deep structure of these networks is constructed by layering these elements, often including input, convolutional, pooling, and flattening layers, as depicted

in Figure 5 [46]. This hierarchical integration of data through these layers culminates in a dense representation, ultimately leading to an output layer that signifies the final processing outcome.

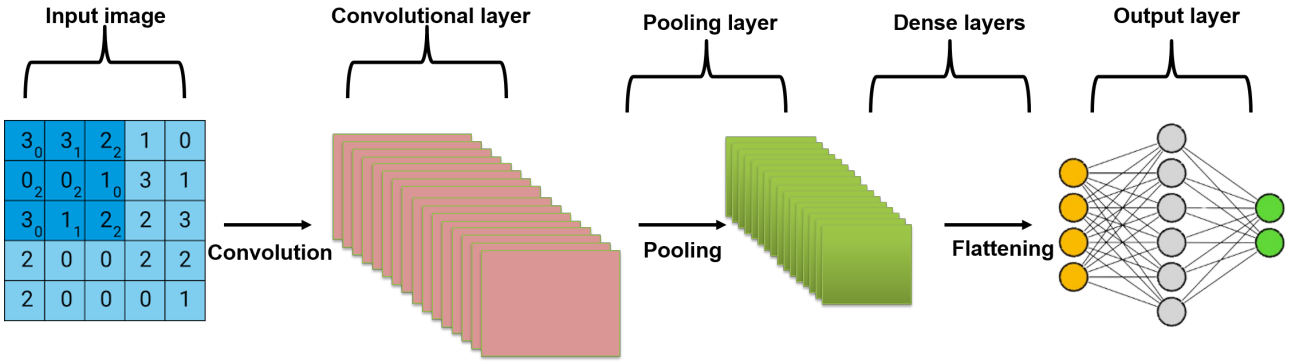

Figure 5. Architecture of CNN method.

A CNN or ConvNet is a deep learning architecture designed to learn directly from data. CNNs are specifically well-suited for identifying patterns in images to recognize objects, classes, and categories. Moreover, they prove to be highly effective in classifying audio, time series, and signal data. In dealing with large volumes of training data, CNN often employs a mini-batch approach. Consider a set of activation features $X = [x_1, \ldots, x_M]$ from M training images, where each feature vector x_i has a dimensionality of K, and $[x_1, \ldots, x_M] \in \{-1, 1\}^M$ stores the ground truth labels. A robust classifier employs a boosting technique to make predictions. The classifier $H(\cdot)$ is a weighted sum of weak classifiers $h(\cdot)$, formulated as Equation (8):

$$H(x_1) = \sum_{j=1}^{K} \alpha_j h(x_{ij}, \lambda_j); \tag{8}$$

$$h(x_{ij}, \lambda_j) = \frac{f(x_{ij}, \lambda_j)}{\sqrt{f(x_{ij}, \lambda_j)^2 + \eta^2}}; \tag{9}$$

where $x_{ij} \in x_j$ is the jth activation feature of the ith image. Each feature $h(x_{ij}, \lambda_j)$ is associated with a potential weak classifier, producing outputs in the range $(-1, 1)$ Equation (9). The function $\frac{f(\cdot)}{\sqrt{f(\cdot)^2+\eta^2}}$ simulates a sign method, optimizing gradient descent through derivative computation. In this framework, $h(x_{ij}, \lambda_j) \in K$ represents a decision stump, a one-level decision tree. The parameter η, adjustable by the distribution of $f(\cdot)$ as σ/c, controls the function's ramp. Here, σ is the standard deviation of $f(\cdot)$, and c is a constant. In this work, η is empirically set to σ/v. The weight of the ith weak classifier $\alpha_j \geq 0$ follows the constraint Equation (10):

$$\sum_{j=1}^{K} \alpha_j = 1; \tag{10}$$

When $\alpha_j = 0$, the corresponding neuron remains inactive in the neural network. The balance between the losses of strong and weak classifiers is given by Equation (11):

$$\epsilon^B = \beta \epsilon^B_{strong} + (1 - \beta)\epsilon_{weak}; \tag{11}$$

where $\beta \in [0, 1]$. The losses for strong and weak classifiers are defined as Equations (12) and (13):

$$\epsilon^B_{strong} = \frac{1}{N} \sum_{i=1}^{N} (H(x_i) - x_i)^2; \tag{12}$$

$$\epsilon_{weak} = \frac{1}{MN} \sum_{i=1}^{N} \sum_{1 \le j \le K, \alpha_j = 0} [h(x_{ij}, \lambda_j) - y_i]^2; \tag{13}$$

where $\alpha_j > 0$ ensures that dormant neurons are excluded from the loss calculation.

3.3. Classification Using CA-WGNN

The Cumulative Attribute-Weighted Graph Neural Network (CA-WGNN) model, which has been used in geometric deep learning, uses a weighted graph structure to map interrelationships between modalities. This groundbreaking approach not only fosters a comprehensive understanding of emotions by tapping into the synergistic potential of diverse data sources but also strives to classify emotions holistically. In the realm of emotion classification, it is paramount to adopt a holistic approach that integrates insights from multiple modalities. This method is especially effective in capturing the nuanced spectrum of human emotions. Specifically, the classification model distinguishes between three primary categories of emotions:

1. **Positive Emotions:** These include feelings such as joy, excitement, and a general sense of positivity. These emotions are characterized by their uplifting and affirmative nature.
2. **Negative Emotions:** This category encompasses emotions like sadness, fear, and disgust, which are typically associated with adverse experiences or perceptions.
3. **Neutral Emotion:** Representing a state that is neither explicitly positive nor negative, this category captures the subtlety of emotions that do not fall distinctly into the other two classifications.

The sophisticated structure of CA-WGNN, which is connectivity-driven, excels in learning from input graphs and deciphering complex system interactions through iterative data analysis across connections. This enables the model to generalize effectively to unfamiliar graphs, making predictions at both the network and node levels.

Thanks to its advanced capabilities, CA-WGNN is instrumental in a variety of applications, showcasing its versatility and effectiveness. In terms of implementation, the pseudocode for CA-WGNN inference, as illustrated in Algorithm 1, delineates the step-by-step process of this model's operation. Moreover, the model's detailed structure, as shown in Figure 6, further elucidates its functional dynamics and connectivity-driven approach, highlighting its role as a pivotal tool in the realm of geometric deep learning.

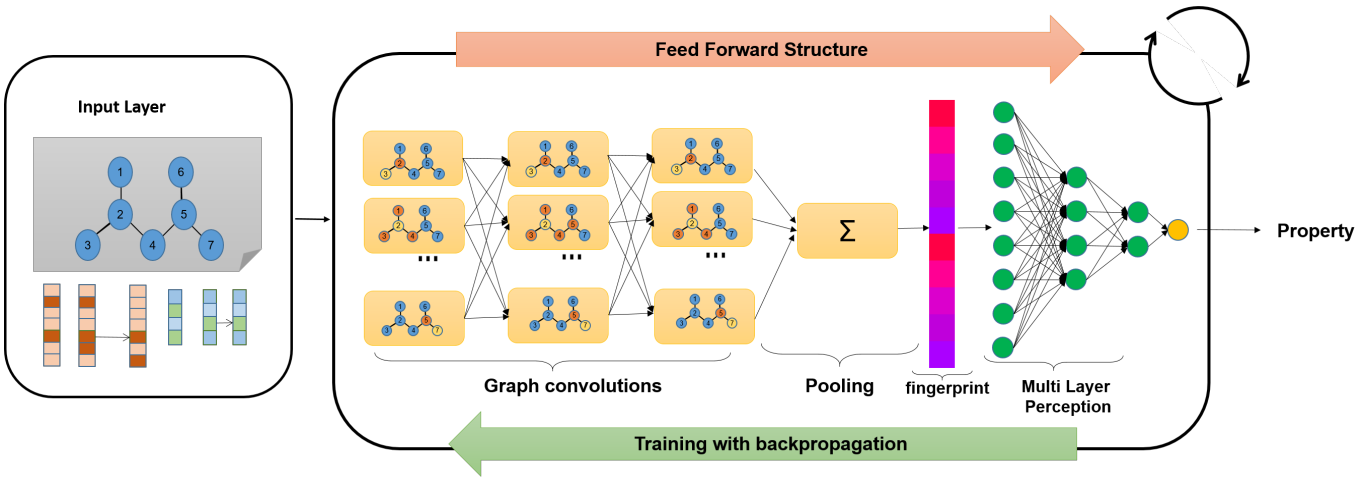

Figure 6. Structure of CA-WGNN.

Algorithm 1 Algorithm of CA-WGNN for emotion recoginition

1: procedure GNN Operator
2: $F \leftarrow$ Number of layers in the GNN
3: $C \leftarrow$ Set of nodes in graph S (assumed static)
4: $A \leftarrow$ Set of edges in graph S (assumed static)
5: Initialize Nodes and Edges:
6: for $c \in C$ do
7: $z_c^0 \leftarrow [y_c, 0, \ldots, 0]$
8: for $a \in A$ do
9: $s_a^0 \leftarrow [h_c, 0, \ldots, 0]$
10: GNN Layered processing:
11: for $f = 1$ to F do
12: Edge processing:*(Order of edge and node processing may be interchanged or even interspersed.)*
13: for $a \in A$ do *(Order of aggregation and combination may be interchanged if aggregation is linear.)*
14: $p_a^{(f)} = \rho_A^{(f)}(s_a^{(f-1)}, z_w^{(f-1)} : w \in M(a))$
15: $s_a^{(f)} = \phi_A^{(f)}(p_a^{(f)})$,
16: Node processing:*(Order of edge and node processing may be interchanged or even interspersed.)*
17: for $c \in C$ do *(The order of aggregation and combination may be interchanged if aggregation is linear.)*
18: $e_c^{(f)} = \rho_C^{(f)}(z_c^{(f-1)}, z_w^{(f-1)} : w \in M(c))$
19: $z_c^{(f)} = \phi_C^{(f)}(e_c^{(f)})$,
20: Readout:
21: $\hat{x} = \phi_S(\rho S(\{z_c^F, s_a^F : c, a \in S\}))$.
22: end procedure.

The CA-WGNN model starts with the graph's edges, vertices, and feature vectors as inputs (step 9). It executes in layers, with each layer updating the feature vectors of all connected vertices and edges concurrently (steps 11, 14, 15). Combination functions, which can be represented by neural networks, are used to update the combined edges and vertices (steps 13, 17). The readout process is carried out by an appropriate function, often a neural network (step 18). The following equations explain the aggregation Equation (14) and combination processes Equation (15) at any given layer $f \in [1, F]$:

$$\text{Aggregation: } P_A^{(f)} = \rho_A^{(f)}(s_a^{(f-1)}, s_w^{(f-1)} : w \epsilon M(a)), \tag{14}$$

$$\text{Combination: } S_A^{(f)} = \phi_A^{(f)}(p_a^{(f)}), \tag{15}$$

where for each layer $f - 1$, the aggregated edge ρ_A includes the edge's feature vector Equation (16) s_a and the terminal vertex feature vectors s_w, where $w \epsilon M(a)$. The combination Equation (17) ϕ_A uses this aggregation as input. Similarly, for nodes:

$$\text{Aggregation: } e_c^{(f)} = \rho_C^{(f)}(z_c^{(f-1)}, z_c^{(f-1)} : w \epsilon M(c)), \tag{16}$$

$$\text{Combination: } z_c^{(f)} = \rho_C^{(f)}(e_c^{(f)}), \tag{17}$$

The final output feature vector $\hat{x}$ is obtained by applying a readout function Equation (18):

$$\text{Readout: } \hat{x} = \phi_S(\rho_S(z_c^F, s_a^F : c, a \in S)) \tag{18}$$

In the domain of speaker characteristic extraction, three distinct methodologies are employed: textual speaker embedding via LSTM networks, acoustic speaker embedding through DNN, and visual speaker embedding utilizing s CNN. The process of emotion

recognition benefits substantially from the amalgamation of these diverse modalities, facilitated by a Cross-Attention Weighted Graph Neural Network CA-WGNN. The DNN effectively models acoustic patterns derived from audio inputs, whereas the LSTM network is adept at discerning temporal relationships in textual data. Concurrently, the CNN excels at extracting spatial features from video frames. Fusion of these modalities is achieved through the construction of a weighted graph, where the edges represent inter-modal correlations and weights are assigned based on the relative importance of each modality. The CA-WGNN leverages these weights to integrate data, enabling the detection of nuanced emotional cues.

This integrative approach fosters a robust and versatile framework for emotion recognition across multiple modalities, thereby enhancing classification accuracy by harnessing complementary information from varied sources. Table 1 delineates the performance of Emotion Recognition Models on the CMU-MOSI andIEMOCAP datasets, detailing their respective experimental configurations.

Table 1. Comparison of Experimental Configurations.

Dataset Name	CMU-MOSI	IEMOCAP
Model Name	CA-WGNN	CA-WGNN
No. of Layers	7	7
Batch Size	32	32
Epochs	100	100
Random State	42	42
Learning Rate	0.001	0.001
Validation Split	0.2	0.5

3.4. Problem Statement

Robust and automated recognition of facial emotion is one of the stated problems. Other issues involve completing meaningful analysis of the taken image based on facial emotions, constructing test and training datasets, and ultimately, developing and implementing fitted classifiers to learn the underlying classifiers' vectors of facial descriptors and deliver a model design that can identify up to six models that are widely accepted across different cultures. Emotions that predominate include dread, joy, sorrow, surprise, disgust, and delight. If the system were in place, it could analyze a person's face and its features to determine their identification based on a weighted assumption. Accurate Sentiment Recognition and identification are made possible by our proposed CA-WGNN method, which combines facial expression analysis, culturally adaptive classifiers, and diverse dataset creation. It is achieved by using cumulative attribute weights, which provide nuanced and cross-cultural accuracy. When faced with the difficult task of automated and robust face emotion identification, the CA-WGNN is an essential tool. When it comes to analyzing the expressed emotions in photographs, CA-WGNN is crucial for solving the given challenge. Its relevance goes beyond only building fitted classifiers to include building complete training and test datasets. For a model to be able to recognize six commonly recognized emotions dread, pleasure, sadness, surprise, disgust, and delight across different cultures, this sophisticated neural network is crucial for learning the complex vectors of face descriptors. The proposed system is enhanced using CA-WGNN so that it can evaluate face characteristics and make reliable identifications using weighted assumptions.

3.5. Assessment Metrics for Classifier Performance

The performance of classifiers is evaluated using key metrics: Accuracy, Macro Precision, Macro Recall, and the F1 score. These metrics are defined as follows:

- **Accuracy**: This metric quantifies the proportion of correctly predicted instances in the dataset and is calculated using the Formula (19):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total}} \tag{19}$$

where TP (True Positives) and TN (True Negatives) are the numbers of correctly identified positive and negative instances, respectively, and Total represents the sum of all instances.

- **Precision**: Defined as the ratio of correctly predicted positive instances to the total predicted as positive, Precision indicates the model's ability to avoid false positives as (20):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{20}$$

FP (False Positives) denotes instances incorrectly classified as positive.

- **Recall**: Recall, or sensitivity, measures the proportion of actual positives that are correctly identified and is given by (21):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{21}$$

FN (False Negatives) are actual positives that were incorrectly classified as negative.

- **F1 Score**: The F1 score is the harmonic mean of Precision and Recall, providing a balance between them, especially important in the presence of class imbalances as (22):

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

This process involves collecting and merging feature vectors from every vertex and edge in the graph at the final iteration F. In this study, distinct methodologies were applied to each dataset to implement our algorithm. For audio data, DNN is utilized to generate speaker embeddings. This involves extracting features from audio files using the librosa library and then processing them through a multi-layer DNN to obtain embeddings. In contrast, for video data, a CNN is employed for a similar purpose. This process includes the extraction and preprocessing of frames from video files, which are then input into the CNN to derive embeddings. The core aspect of this research lies in the classification of these embeddings from each modality. In all three datasets text, audio, and video embeddings are separated into features (X) and target values (y). These are then input into a specially designed Cumulative Attribute Weighted Graph Neural Network (CA-WGNN) model. The model's performance is evaluated using metrics such as accuracy, precision, recall, and F1 score. These metrics collectively offer a comprehensive view of the classifier's performance, taking into account both error types and class distribution in the dataset.

4. Experimental Results

In this study, the experimental framework was meticulously designed to accommodate the rigors of machine learning applications, leveraging a harmonious blend of hardware and software configurations. Central to this framework was a Windows 10 operating system, chosen for its robustness and compatibility with advanced computational tasks. The experimental environment encompassed a dual-platform approach, utilizing both Google Colab and Jupyter Notebook. This methodology facilitated a comprehensive evaluation, allowing for a detailed analysis of various challenges, as enumerated in the accompanying Table 2. A notable aspect of this setup was the dependency of Google Colab's time efficiency on internet connectivity and hardware specifications. From a hardware perspective, the experiments were conducted on a system equipped with a Windows 10 64-bit operating system, complemented by 200 GB of storage in the C drive and 8 GB of RAM. This configuration ensured a seamless operational flow during the testing phases. Python, executed via Jupyter Notebook, was the primary programming language employed, underscoring its versatility and efficiency in handling complex datasets. Integral to our research was the application of GNNs for sentiment analysis, targeting multimodal

data encompassing text, audio, and video. This innovative approach, utilizing GNNs, was pivotal in analyzing the interconnected elements within these data types, thereby enhancing the accuracy and depth of sentiment analysis. The incorporation of GNNs allowed for a more nuanced understanding of the sentiment conveyed across different modalities, significantly augmenting the scope of our study in the domain of TER.

The CA-WGNN method is introduced for emotion recognition, leveraging trimodal inputs, including text, audio, and video modalities. This approach aims to enhance the interaction among emotional inputs, thereby fostering a more sophisticated and efficient system. The CA-WGNN is evaluated against established methodologies, such as the Support Vector Machine (SVM) [47], the Voting Classifier (integrating Logistic Regression and Stochastic Gradient Descent, denoted as the Voting Classifier, Logistic Regression and Stochastic Gradient Descent (VC(LR-SGD))) [47], and the Gradient Boosting Model (GBM) [47]. Metrics, including accuracy, precision, recall, and the F1-score, are employed to determine the efficacy of inaccurate emotion categorization models. These metrics are essential for assessing the models' robustness across various applications, confirming their reliability and capability to discern complex emotional nuances. In their comprehensive study, Yousaf A. et al. [47] conducted a thorough evaluation of the Support Vector Machine (SVM) alongside six other prominent machine learning algorithms: Logistic Regression, Stochastic Gradient Descent, Naive Bayes, Decision Tree, Random Forest, and Voting Classifier, in recognizing emotions through textual tweet classification. This investigation focused on assessing and comparing the performance metrics of these models. Notably, a linear kernel function was employed for the SVM, indicating that the decision boundary is linear. The transformation of tweets into numerical vectors was facilitated through the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method, which emphasizes the significance and frequency of each word within the tweets. The study adhered to the default parameters for various SVM configurations, including the regularization parameter 'C', which balances the margin width and classification error, the kernel coefficient 'gamma', which affects the influence of individual training samples, the tolerance for the stopping criterion 'tol', and the maximum number of iterations 'maxiter.' The research findings demonstrated that SVM was significantly effective, achieving an accuracy of 77% and an F1-score of 79% in a binary dataset that categorized tweets as happy or unhappy. Furthermore, in a multi-class dataset, where tweets were classified as happy, sad, angry, or neutral, SVM attained an accuracy of 67% and an F1 score of 64%. This in-depth analysis underscores the SVM's capability to efficiently recognize emotions from textual content on social media platforms, particularly Twitter, providing valuable contributions to sentiment analysis and emotion detection.

Table 2. Hardware and software configuration for system requirements.

Component	Specification
RAM	8 GB
Hard Disk	200 GB in C Drive
OS	Windows 10 64 bit
Software	Individual Edition-Anaconda Jupyter Notebook
Source	https://www.anaconda.com/products/individual (accessed on 26 February 2024) https://docs.anaconda.com/anaconda/install/windows/ (accessed on 26 February 2024) https://jupyter.org (accessed on 26 February 2024)
Total Processing Time	(h, m, s): 00, 21, 31

4.1. Accuracy Analysis

Among the multiple factors considered in emotion detection, accuracy stands as a paramount measure, particularly when comparing the performance of models across multimodal datasets like IEMOCAP Table 3 and CMU-MOSI Table 4. This comparative study focuses on the accuracy in recognizing emotions through text, audio, and video modalities. It delves into the model's proficiency in identifying and classifying emotional expressions accurately across these varied channels. The essence of this analysis is to gauge how effectively the trimodal method captures and aligns with the diverse emotional nuances present in both datasets. Such an evaluation provides pivotal insights into the model's overall capability to decipher emotions, whether conveyed through textual, auditory, or visual means.

Table 3. Performance Metrics for IEMOCAP Dataset

Dataset	Category	Precision (%)	Recall (%)	F1-Score (%)	Support (%)
IEMOCAP Text	Negative	0.91	0.90	0.91	69
	Neutral	0.96	0.96	0.96	68
	Positive	0.92	0.94	0.93	63
IEMOCAP Audio	Negative	0.91	0.91	0.91	69
	Neutral	0.94	0.96	0.95	68
	Positive	0.94	0.92	0.93	63
IEMOCAP Video	Negative	0.89	0.93	0.91	69
	Neutral	0.96	0.94	0.95	68
	Positive	0.93	0.90	0.92	63
Fusion Overall Accuracy					0.93 (200)
Fusion Macro Average		0.93	0.92	0.93	200
Fusion Weighted Average		0.93	0.93	0.93	200

Table 4. Performance Metrics for CMU-MOSI Datasets.

Dataset	Category	Precision (%)	Recall (%)	F1-Score (%)	Support (%)
CMU-MOSI Text	Negative	0.88	0.94	0.91	69
	Neutral	0.97	0.90	0.93	68
	Positive	0.90	0.90	0.90	63
CMU-MOSI Audio	Negative	0.90	0.93	0.91	69
	Neutral	0.98	0.94	0.96	68
	Positive	0.91	0.92	0.91	63
CMU-MOSI Video	Negative	0.88	0.96	0.92	69
	Neutral	1.00	0.94	0.97	68
	Positive	0.92	0.89	0.90	63
Fusion Overall Accuracy					0.94 (200)
Fusion Macro Average		0.94	0.94	0.94	200
Fusion Weighted Average		0.94	0.94	0.94	200

The results presented in the following sections compare the accuracy of the IEMOCAP and CMU-MOSI datasets, based on the criteria of Sentiment Recognition.

The analysis revealed a marginally higher accuracy in the CMU-MOSI dataset (0.94%) compared to IEMOCAP (0.93%), indicating nuanced differences in the performance of these models in emotion detection across the three modalities. This comparison not only highlights the strengths and limitations of each dataset but also underscores the importance of multimodal approaches in enhancing the accuracy of sentiment recognition systems.

The provided table presents a comprehensive overview of the performance metrics for the CMU-MOSI dataset, categorized into three modalities: Text, Video, and Audio. Each

modality is further subdivided into Negative, Neutral, and Positive sentiment categories. The table displays essential evaluation metrics, including Precision, Recall, and F1-Score, which are crucial for assessing the effectiveness of a sentiment classification model. Notably, the overall accuracy of the model across all modalities is 94%, based on a total of 200 data points (Figure 7). The "Macro Average" and "Weighted Average" rows summarize the performance across categories, showcasing the model's ability to maintain high precision, recall, and F1 scores across the diverse CMU-MOSI dataset.

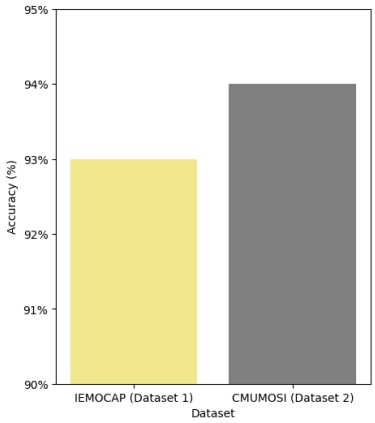

Figure 7. The accuracy of the IEMOCAP and CMU-MOSI datasets.

Following the analytical discourse, we present Table 5, which succinctly compares the accuracy percentages obtained from the IEMOCAP and CMU-MOSI datasets. This table serves as a quantitative testament to the findings discussed above, illustrating the slight yet significant variance in the accuracy of Sentiment Recognition between the two datasets.

Table 5. Comparison of accuracy between the IEMOCAP and CMU-MOSI datasets.

Dataset	Accuracy (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

The data presented in Table 5 provides a clear and concise reference for understanding the efficacy of each dataset in the context of TER. It reinforces the notion that while both datasets exhibit high levels of accuracy, subtle distinctions in their performance can be critical in selecting the appropriate dataset for specific research or application needs.

Across the field of emotion recognition, accurately determining the effectiveness of various computational methods is crucial for advancing the field. Table 6 and Figure 8 present the accuracy of both the proposed and existing methods. The CA-WGNN achieved 94%, compared with the existing methods of SVM, which attained 76%, GBM attained 74%, and VC(LR-SGD) attained 79%. It demonstrates that the suggested process achieves higher rates than existing methods, underscoring the superiority of the CA-WGNN approach in identifying emotions with greater accuracy. This notable improvement highlights the CA-WGNN in enhancing the accuracy and reliability of emotion recognition systems, marking a significant step forward in the field.

Table 6. Numerical outcomes of Accuracy.

Methods	Accuracy (%)
SVM [47]	76
GBM [47]	74
VC(LR-SGD) [47]	79
CA-WGNN [Proposed]	94

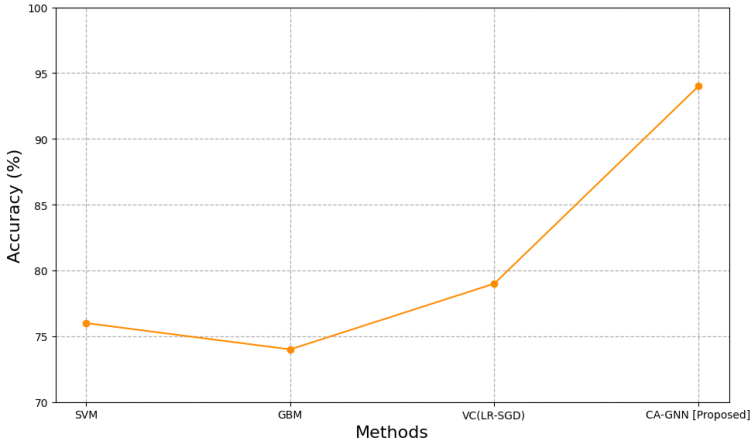

Figure 8. Graphical representation of Accuracy.

4.2. Precision Analysis

Delving into the intricacies of trimodal emotion detection, this study extends its focus to the precision aspect in the evaluation of text, audio, and video components within the IEMOCAP and CMU-MOSI datasets. Precision in this context is a measure of how accurately the trimodal emotion identification model discerns and categorizes emotions, taking into account the variances in emotional expressions manifesting through visual, auditory, and textual forms.

A crucial part of this comparative analysis involves assessing the diversity, quality, and annotation methodologies of the datasets. These elements are instrumental in understanding the nuances and the resulting precision in sentiment recognition.

The detailerd comparison of precision levels achieved by the models using IEMOCAP and CMU-MOSI datasets is systematically presented in Table 7 and Figure 9. The findings indicate a marginally superior precision in the CMU-MOSI dataset (0.94%) as compared to IEMOCAP (0.93%).

Table 7. The comparison of precision between the IEMOCAP and CMU-MOSI datasets.

Dataset	Precision (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

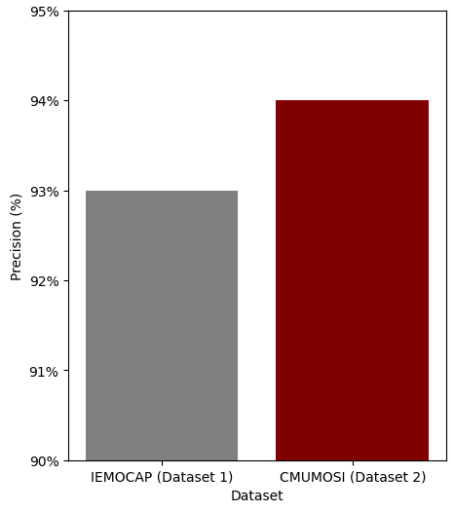

Figure 9. The comparison of precision between the IEMOCAP and CMU-MOSI datasets.

This precision comparison not only highlights the subtle yet significant differences in the datasets but also underscores the importance of detailed and nuanced evaluation in TERsystems. An analysis such as this is vital to developing more accurate and reliable sentiment recognition models.

The comparative analysis of precision metrics significantly highlights the effectiveness of different computational approaches in emotion recognition. In Table 8 and Figure 10 we present the precision for both the proposed and existing methods. The CA-WGNN achieved a remarkable 92%, in comparison with the existing methods of SVM, which attained 76%, GBM attained 72%, and VC(LR-SGD) attained 78%. This demonstrates that the suggested process achieves higher precision rates than the existing methods, underlining the superiority of the CA-WGNNapproach in accurately categorizing emotional states. The results reinforce the notion that leveraging advanced algorithms like CA-WGNN significantly enhances the precision of emotion recognition, setting a new benchmark for future research in this domain.

Table 8. Numerical Outcomes of Precision.

Methods	Precision (%)
SVM [47]	76
GBM [47]	72
VC(LR-SGD) [47]	78
CA-WGNN [Proposed]	92

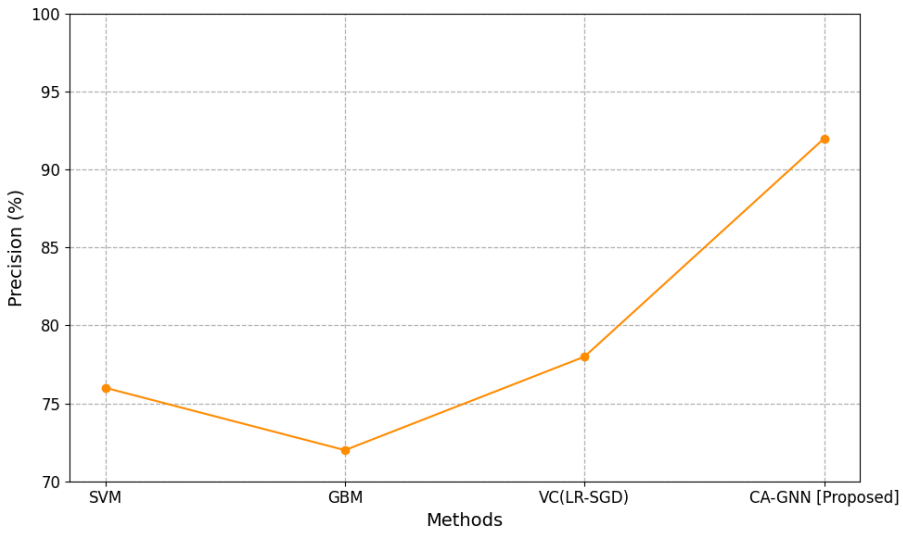

Figure 10. Graphical representation of Precision.

4.3. Recall Analysis

In the intricate landscape of Sentiment Recognition, recall emerges as a critical metric, reflecting the model's capacity to accurately identify a wide range of emotions. This study delves into a comparative analysis of recall within the IEMOCAP and CMU-MOSI datasets, each known for their unique strengths and focuses. IEMOCAP excels in capturing a broad spectrum of realistic emotional expressions in audiovisual data, whereas CMU-MOSI distinguishes itself with an emphasis on multilingual Sentiment Recognition. The application of a tri-modal emotion identification approach in this study is pivotal. It facilitates a more comprehensive and context-aware categorization of emotions, adaptable to a variety of situations. This methodology enhances the depth and breadth of Sentiment Recognition, moving beyond conventional single-modal approaches.

Table 9 and Figure 11 present a side-by-side comparison of the recall capabilities of models trained on these datasets. The results indicate a slightly higher recall for the

CMU-MOSI dataset (0.94%) compared to IEMOCAP (0.93%), underscoring the nuanced differences in the datasets' abilities to accurately recognize a diverse range of emotions.

Table 9. Comparison of Recall.

Dataset	Recall (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

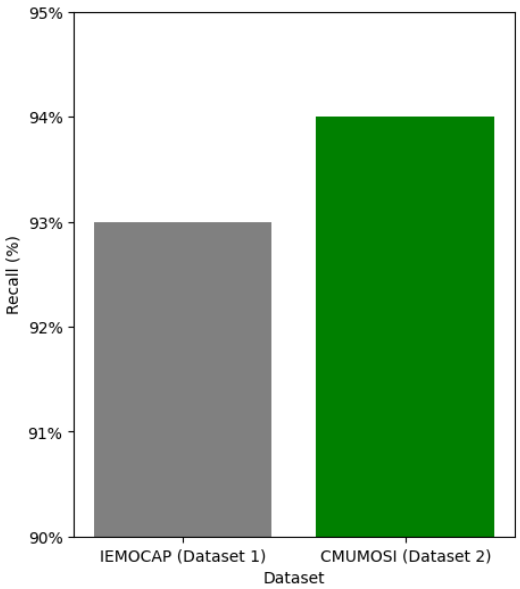

Figure 11. The recall comparison between the IEMOCAP and CMU-MOSI datasets.

This comparison not only highlights the specific strengths of each dataset but also emphasizes the importance of employing multimodal Sentiment Recognition frameworks for more accurate and versatile ER systems. The findings bolster the case for the ongoing refinement and enhancement of these technologies in the field of ER research.

The evaluation of precision across various methods is critical in understanding their effectiveness in emotion recognition. Table 10 and Figure 12 present the precision metrics for both the proposed and existing methods. The CA-WGNN method demonstrates superior performance by achieving a 93% precision rate, in comparison with the existing methods where SVM attained 80%, GBM attained 79%, and VC(LR-SGD) attained 84%. This comparison clearly illustrates that the suggested CA-WGNN process outperforms the existing methods in terms of precision. The findings underscore the CA-WGNN's advanced capability in accurately identifying emotional nuances, setting a new benchmark for future research in the domain of emotion recognition.

Table 10. Numerical Outcomes of Recall.

Methods	Recall (%)
SVM [47]	80
GBM [47]	79
VC(LR-SGD) [47]	84
CA-GNN [Proposed]	93

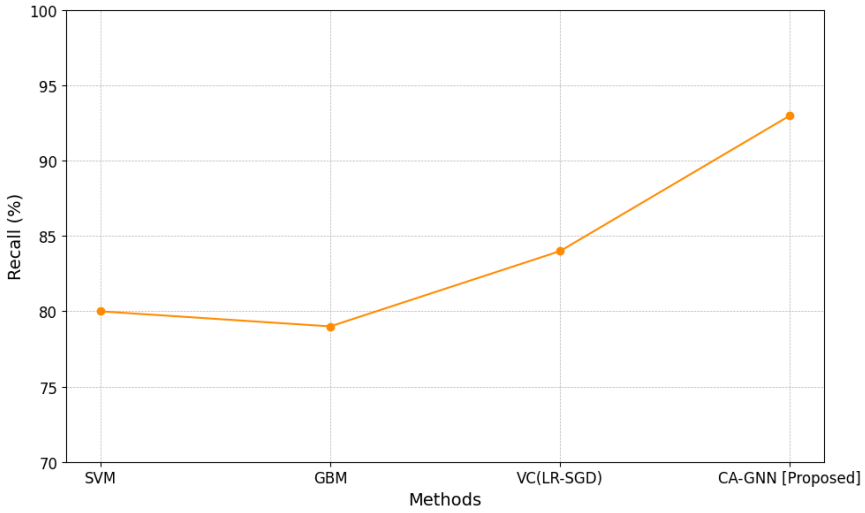

Figure 12. Graphical representation of Recall.

4.4. F1-Score Analysis

The F1-score, a crucial statistic in Sentiment Recognition systems, serves as a harmonious balance between precision and recall, encapsulating the overall accuracy of the system. In the context of trimodal emotion detection, the F1-score acquires an elevated significance as it reflects the system's ability to effectively integrate and interpret data across text, audio, and video modalities.

In this study, the robustness and adaptability of the trimodal emotion detection technique are scrutinized by comparing the F1-scores obtained from the IEMOCAP and CMU-MOSI datasets. Achieving high F1 scores is indicative of a model's precision in identifying emotions, thereby underscoring the system's efficiency across diverse datasets and modalities. This is crucial for gaining a comprehensive understanding of emotional expressions.

Table 11 and Figure 13 present a comparative analysis of the F1-scores for the IEMOCAP and CMU-MOSI datasets, based on their performance in ER. The data reveals a marginally higher F1-score for the CMU-MOSI dataset (0.94%) as compared to that of IEMOCAP (0.93%), highlighting the nuanced efficacy of the models in processing emotional data.

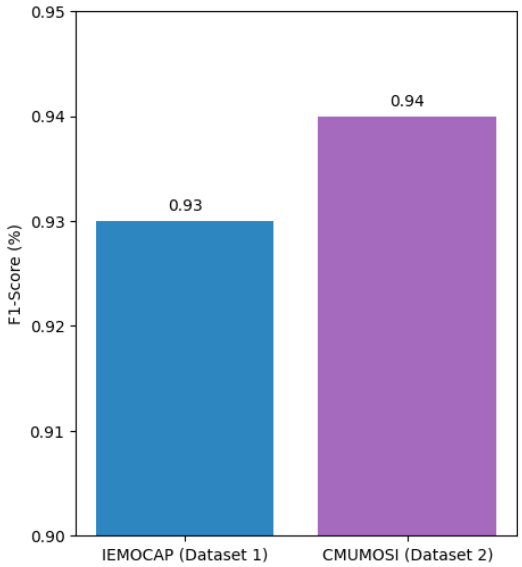

Figure 13. The F1-scores comparison between the IEMOCAP and CMU-MOSI datasets.

Table 11. Comparison of F1-score.

Dataset	F1-Score (%)
IEMOCAP (Dataset 1)	0.93
CMU-MOSI (Dataset 2)	0.94

This comparative exploration of F1 scores not only sheds light on the performance metrics of the datasets but also provides invaluable insights for future advancements in TERsystems, aiming for models that are both accurate and versatile.

The meticulous comparison of precision metrics underscores the advancements and effectiveness of emotion recognition technologies. In Table 12 and Figure 14, we detail the precision rates for both the proposed and existing methods. The CA-WGNN method stands out by achieving a remarkable 91% precision, in contrast to the existing methods where SVM and GBM each attained 76%, and VC(LR-SGD) reached 81%. This comparison vividly illustrates the superior performance of the proposed CA-WGNN method, affirming its capacity to achieve higher precision rates than the conventional approaches. The results highlight the efficiency of the CA-WGNN and pave the way for future innovations in the field of emotion recognition, emphasizing the potential for further advancements and the continuous need for research and development in this dynamic area.

Table 12. Numerical Outcomes of F1-score.

Methods	F1-Score (%)
SVM [47]	78
GBM [47]	76
VC(LR-SGD) [47]	81
CA-GNN [Proposed]	91

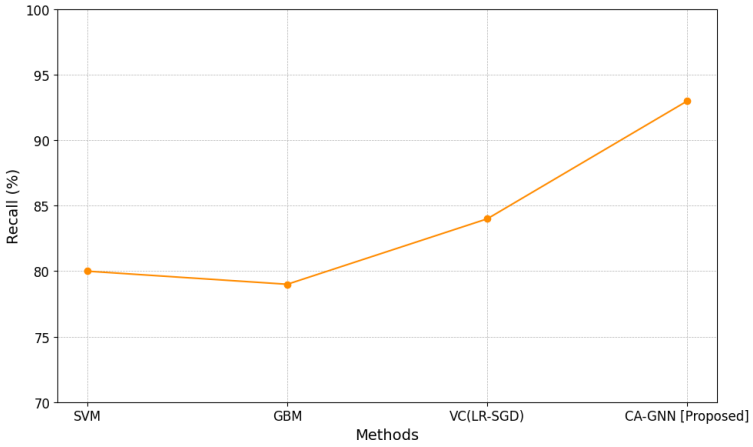

Figure 14. Graphical representation of F1-score.

4.5. Advancements Brought by CA-WGNN

The diagram depicts the architecture of the CA-WGNN model, highlighting the sequential data flow through various layers, each contributing significantly to the model's classification capabilities.

- **Cumulative Attribute Weighting:** The CA-WGNN model introduces a novel cumulative weighting approach for attributes within a graph structure. This technique is instrumental in deepening the understanding of node interrelations and dependencies within the graph.
- **Enhanced Contextual Understanding:** By focusing on cumulative attributes, CA-WGNNs achieve a heightened comprehension of the broader context surrounding each node. This aspect is crucial for making accurate predictions in complex graph structures.

- **Applicability in Complex Network Analysis:** CA-WGNNs are particularly advantageous in scenarios involving intricate graph structures. In such contexts, a node's influence often extends beyond its immediate neighbors, necessitating a more sophisticated analysis approach.

The effectiveness of the CA-WGNN model is further evidenced by its performance metrics on the IEMOCAP dataset, as shown in the Table 13. The model demonstrates high precision, recall, and F1-score across different categories, underscoring its robustness and accuracy in sentiment analysis. Notably, the model achieves an overall accuracy of 94%, with a balanced performance across all categories, as indicated by the macro and weighted averages. These results highlight the model's capability in accurately classifying sentiments, even in the challenging context of multimodal datasets.

Table 13. Performance metrics of the CA-WGNN model on the IEMOCAP dataset.

Category	Precision	Recall	F1-Score	Support
Negative	0.86	0.93	0.90	69
Neutral	0.96	0.97	0.96	68
Positive	0.96	0.87	0.92	63
Accuracy			0.93 (200)	
Macro Average		0.93	0.92	0.93
Weighted Average		0.93	0.93	0.93

The CA-WGNN model's performance was rigorously assessed using the CMU-MOSI dataset (Table 14), known for its comprehensive and multimodal nature. This evaluation aimed to ascertain the model's proficiency in accurately analyzing and classifying sentiments across diverse categories. The table below presents a detailed breakdown of the model's performance metrics, including precision, recall, and F1-score for Negative, Neutral, and Positive sentiment categories, along with the overall accuracy, macro, and weighted averages.

Table 14. Performance Metrics of the CA-WGNN Model on the CMU-MOSI Dataset.

Category	Precision (%)	Recall (%)	F1-Score (%)	Support (%)
Negative	0.93	0.94	0.94	69
Neutral	0.97	0.94	0.96	68
Positive	0.92	0.94	0.93	63
Accuracy			0.94 (200)	
Macro Average		0.94	0.94	0.94
Weighted Average		0.94	0.94	0.94

The data in the table reveal the CA-WGNN model's exceptional accuracy and consistency in sentiment classification. The model demonstrates high precision and recall rates across all categories, a testament to its robust algorithmic structure. The precision of 0.97 in the Neutral category underscores the model's capability to discern nuanced emotional expressions, a challenging aspect in affective computing. The overall accuracy of 94% is indicative of the model's effectiveness in dealing with the intricacies of multimodal data, as presented in the CMU-MOSI dataset. These results validate the CA-WGNN model's theoretical underpinnings and establish its practical applicability in real-world scenarios, particularly in the domain of multimodal affective computing. The comprehensive performance analysis using the CMU-MOSI dataset offers valuable insights into the model's potential for broader applications in sentiment analysis and emotional recognition. The table above provides a quantitative evaluation of the CA-WGNN model's performance. The reported metrics, namely precision, recall, and F1-score, reflect the model's effectiveness in various sentiment categories (Negative, Neutral, and Positive).

With an overall accuracy of 94%, the CA-WGNN demonstrates its robustness in analyzing and classifying complex network structures. Notably, its superior performance in the Neutral category, with a precision of 98% and an F1-score of 96%, underscores the model's nuanced understanding of subtle contextual variations. These results validate the efficacy of the cumulative attribute weighting and enhanced contextual understanding principles inherent in the CA-WGNN design.

5. Discussion

In the evolving field of emotion recognition, the IEMOCAP and CMU-MOSI datasets are essential, encompassing diverse text, audio, and video modalities. Their extensive multimodal data are crucial for those studying emotional expression. These datasets' varied participant demographics, emotional contexts, and data collection methods impact the applicability of models across different datasets, highlighting challenges in effectively merging data from these varied modalities. Traditional approaches often fall short of understanding the complex interplay among different modalities.

Exploring the landscape of emotional recognition technologies, we encounter several challenges across different models. Analyzing complex emotional data with SVM is difficult due to its sensitivity to kernel, parameter choices, and computational demands. Likewise, the VC(LR-SGD) faces hurdles, including hyperparameter sensitivity, potential biases in unbalanced datasets, and the presupposition that all factors contribute equally. The GBM also grapples with issues like overfitting, handling unbalanced datasets, and significant computational expenses. In conclusion, while each model offers potential for emotional recognition, their limitations underscore the need for careful consideration in model selection and the ongoing pursuit of more adaptable and efficient solutions.

This research introduces the Cumulative Attribute- Weighted Graph Neural Network CA-WGNN model, an innovative breakthrough in Sentiment Recognition. The CA-WGNN's graph-based structure excels in attributing weights to features across modalities, adeptly recognizing inter-modal connections, and assigning relevant weights to each modality's features. Its versatility with different modalities and specific dataset characteristics, like those of IEMOCAP and CMU-MOSI, showcases its capability in universalizing Sentiment Recognition models across various datasets. The CA-WGNN's sophisticated grasp of intricate relationships and effective feature integration marks notable progress in this field. The study also offers a detailed comparison of experimental outcomes in subsequent Tables, discussing the CA-WGNN algorithm's benefits, challenges, and environmental impacts. Table 15 presents the accuracy, precision, recall, and F1-scores from IEMOCAP and CMU-MOSI, confirming the algorithm's proficiency in Sentiment Recognition. Table 16 outlines the key benefits and obstacles in implementing CA-WGNN, along with its environmental effects in real-world applications. These tables collectively illuminate the algorithm's performance and flexibility in various environments, underscoring the importance of multimodal integration in affective computing. This comparison particularly emphasizes CA-WGNN's skill in managing complex, cross-modal interactions, vital to the evolution of affective computing. In the context of emotion detection in the text, particularly when examining datasets such as CMU-MOSI, which encompass textual, audio, and visual data, the Receiver Operating Characteristic (ROC) as shown in the Figure 15 curve emerges as a crucial analytical tool. If we take a closer look at the ROC curve using examples from the text portion of two distinct CMU-MOSI datasets, we can gain insights into the model's capability to differentiate between various emotional states expressed in the data.

Figure 15 curve graphically illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different thresholds. By analyzing the happiness and sadness, present in the text data. The area under the ROC curve, we can evaluate how effectively the model distinguishes between different emotional states, such as happiness and sadness, present in the text data. The area under the ROC curve Area Under the Curve (AUC) offers a concise measure of the model's overall performance, with a higher AUC indicating greater accuracy. This evaluation is particularly vital in scenarios

where maintaining a balance between correctly identifying emotions (true positives) and minimizing false identifications (false positives) is paramount. Thus, the ROC curve serves as an indispensable benchmark in the optimization of emotion detection models, ensuring they accurately and reliably interpret the complex array of human emotions embedded in text, as well as in multimodal data like that of the AUC datasets.

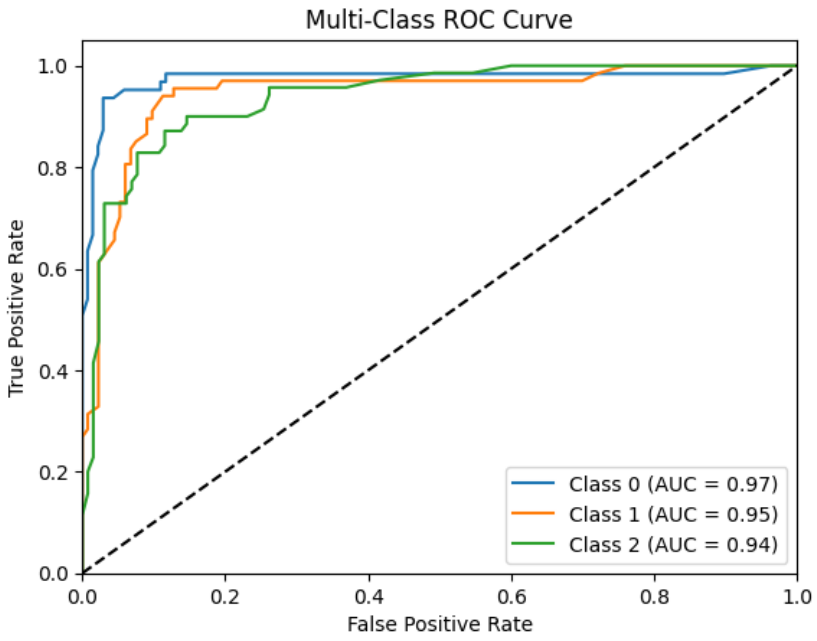

Figure 15. ROC Curve Analysis for Text Data from the CMU-MOSI Dataset.

The CA-WGNN model offers numerous advantages in the Graph Neural Networks (GNNs) domain for emotion recognition and sentiment analysis. Its graph-based framework enables a deeper insight into the interconnected nature of multimodal data, effectively modeling the complex dynamics of emotional expression for more accurate sentiment analysis. The model's capacity to apply different weights to various modalities allows for customized approaches for each dataset, ensuring effectiveness across datasets with diverse structures and content. Moreover, CA-WGNN's focus on cumulative weighting enhances emotion recognition accuracy. By prioritizing impactful features in an emotional context, the model ensures that more relevant data are emphasized, improving emotion and sentiment detection precision while reducing misinterpretation from less pertinent or noisy data.

Overall, the CA-WGNN model signifies a substantial advancement in GNN applications for emotion recognition and sentiment analysis. Its cutting-edge structure and strategic weighting mechanism adeptly capture human emotional nuances, leading to more dependable and insightful affective computing analyses.

Table 15. Comparison of the result.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
IEMOCAP (Dataset 1)	0.93	0.93	0.93	0.93
CMU-MOSI (Dataset 2)	0.94	0.94	0.94	0.94

Table 16. Challenges of CA-WGNN.

Advantages	Challenges	Environmental Effects
Through effective text, audio, and video integration, CA-WGNN captures complex interactions for comprehensive ER. CA-WGNN overcomes standard models' inefficient fusion and captures cross-modal interactions using a graph-based design.	The model's training and inference processes can require more computer resources due to the graph-based design's complexity.	CA-WGNN's enhanced performance and robustness to help multimodal emotion identification systems in real-world applications.
Using two separate datasets, IEMOCAPS and CMU-MOSI, the model shows that it can generalize to varied emotional circumstances and demographics of participants.	Although CA-WGNN is more adaptable, datasets can remain with different participant demographics and data-collection methods, making generalization difficult.	The model's flexibility and integration of multiple modalities can require more computing resources in resource-constrained contexts.
CA-WGNN uses a cumulative weighting method to improve the model's data interpretation by assigning relative values to features in each modality.	The cumulative weighing method may make it difficult to determine each attribute's decision-making impact.	In varied emotional settings, the model's capacity to generalize across datasets makes it useful.

6. Conclusions

The Trimodel emotion identification system, powered by the innovative CA-WGNN method, marks a significant leap in the realm of Sentiment Recognition. This system excels in harmonizing the unique strengths of textual, audio, and visual modalities, weaving them together through a sophisticated weighted graph structure. Its prowess lies not just in the seamless integration of these diverse data sources but also in its superior performance in detecting nuanced emotional states. This effectiveness is vividly demonstrated in comparative experiments, where it eclipses existing state-of-the-art models. A key aspect of this system is its adept modality-specific feature extraction, complemented by a comprehensive approach to emotional categorization. This dual strategy doesn't just enhance accuracy; it also illuminates the intricate interplay between different emotional expressions and the modalities through which they're perceived. The potential applications of this approach are vast and varied, spanning from mental health assessments to enriching HCI. When benchmarked against two prominent datasets, IEMOCAP and CMU-MOSI, the system showcases its robustness, particularly with the CMU-MOSI dataset. Here, it achieves an impressive accuracy of 0.94%, along with equally remarkable precision, recall, and f1-score, all clocking in at 0.94%. This level of precision underscores the system's ability to offer a deeper, more nuanced understanding of emotions, paving the way for more advanced affective computing applications. Looking ahead, this model is poised to revolutionize the field of affective computing. In an ever-evolving technological landscape, the TER system stands out as a groundbreaking development, pushing the boundaries of emotion detection research. Its introduction heralds a new era in multimodal affective computing methodologies, setting the stage for further explorations and innovations.

Author Contributions: Conceptualization, H.F.T.A.-S. and R.D.; methodology, H.F.T.A.-S. and R.D.; writing—original draft preparation, H.F.T.A.-S.; writing—review and editing, R.D.; supervision, R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Experiments used publicly available datasets: IEMOCAP: https://paperswithcode.com/dataset/iemocap (accessed on 26 February 2024); CMU-MOSI: http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/ (accessed on 26 February 2024).

Acknowledgments: This paper was produced from the doctoral thesis titled "Graph Neural Network Based Multimodal Emotion Recognition" presented at Fırat University, Graduate School of Natural and Applied Sciences, Department of Software Engineering, under the supervision of Resul Daş.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CA-WGNN	Cumulative Attribute-Weighted Graph Neural Network
CMU-MOSI	Carnegie Mellon University Multimodal Opinion-level Sentiment Intensity Dataset
AUC	Area Under the Curve
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Networks
ER	Emotion Recognition
SVM	Support Vector Machine
GBM	Gradient Boosting Model
VC(LR-SGD)	the Voting Classifier, Logistic Regression and Stochastic Gradient Descent
IEMOCAP	Interactive Emotional Dyadic Motion Capture
GNNs	Graph Neural Networks
HCI	Human-Computer Interaction
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
ROC	Receiver Operating Characteristic
SA	Sentiment Analysis
SER	Speech Emotion Recognition
TER	Trimodal Emotion Recognition

References

1. Szymkowiak, A.; Gaczek, P.; Jeganathan, K.; Kulawik, P. The impact of emotions on shopping behavior during an epidemic. What a business can do to protect customers. *J. Consum. Behav.* **2021**, *20*, 48–60. [CrossRef]
2. Pal, S.; Mukhopadhyay, S.; Suryadevara, N. Development and progress in sensors and technologies for human emotion recognition. *Sensors* **2021**, *21*, 5554. [CrossRef]
3. Kosti, R.; Alvarez, J.; Recasens, A.; Lapedriza, A. Context-based emotion recognition using emotic dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2755–2766. [CrossRef]
4. Marmpena, A. Emotional Body Language Synthesis for Humanoid Robots. Ph.D. Thesis, University of Plymouth, Plymouth, UK, 2021. [CrossRef]
5. Sarker, I. Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Comput. Sci.* **2022**, *3*, 158. [CrossRef]
6. Dzedzickis, A.; Kaklauskas, A.; Bucinskas, V. Human emotion recognition: Review of sensors and methods. *Sensors* **2020**, *20*, 592. [CrossRef] [PubMed]
7. Baffour, P.; Nunoo-Mensah, H.; Keelson, E.; Kommey, B. A Survey on Deep Learning Algorithms in Facial Emotion Detection and Recognition. *Inform J. Ilm. Bid. Teknol. Inf. Dan Komun.* **2022**, *7*, 24–32. [CrossRef]
8. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [CrossRef] [PubMed]
9. Hamed, S.; Ab Aziz, M.; Yaakub, M. Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments. *Sensors* **2023**, *23*, 1748. [CrossRef] [PubMed]
10. Khurana, Y.; Gupta, S.; Sathyaraj, R.; Raja, S. RobinNet: A Multimodal Speech Emotion Recognition System with Speaker Recognition for Social Interactions. *IEEE Trans. Comput. Soc. Syst.* **2022**, *11*, 478–487. [CrossRef]
11. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]
12. Karna, M.; Juliet, D.S.; Joy, R. Deep learning based Text Emotion Recognition for Chatbot applications. In Proceedings of the 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), Tirunelveli, India, 15–17 June 2020; pp. 988–993. [CrossRef]
13. Cai, L.; Hu, Y.; Dong, J.; Zhou, S. Audio-Textual Emotion Recognition Based on Improved Neural Networks. *Math. Probl. Eng.* **2019**, *2019*, 2593036. [CrossRef]
14. Chen, K.; Gong, S.; Xiang, T.; Loy, C.C. Cumulative Attribute Space for Age and Crowd Density Estimation. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2467–2474. [CrossRef]

15. Ortega, J.D.S.; Senoussaoui, M.; Granger, E.; Pedersoli, M.; Cardinal, P.; Koerich, A.L. Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition. *arXiv* **2019**, arXiv:1907.03196.
16. Chen, K.; Jia, K.; Huttunen, H.; Matas, J.; Kämäräinen, J.K. Cumulative attribute space regression for head pose estimation and color constancy. *Pattern Recognit.* **2019**, *87*, 29–37. [CrossRef]
17. Savci, P.; Das, B. Comparison of pre-trained language models in terms of carbon emissions, time, and accuracy in multi-label text classification using AutoML. *Heliyon* **2023**, *9*, e15670. [CrossRef]
18. Nie, W.; Yan, Y.; Song, D.; Wang, K. Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition. *Multimed. Tools Appl.* **2021**, *80*, 16205–16214. [CrossRef]
19. Pranav, E.; Kamal, S.; Satheesh Chandran, C.; Supriya, M. Facial Emotion Recognition Using Deep Convolutional Neural Network. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 March 2020; pp. 317–320. . [CrossRef]
20. Dolka, H.; M, A.X.V.; Juliet, S. Speech Emotion Recognition Using ANN on MFCC Features. In Proceedings of the 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, 13–14 May 2021; pp. 431–435. [CrossRef]
21. Huddar, M.; Sannakki, S.; Rajpurohit, V. Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN. *Int. J. Interact. Multimed. Artif. Intell.* **2021**, *7*, 44–51. [CrossRef]
22. Schmitz, M.; Ahmed, R.; Cao, J. Bias and fairness on multimodal emotion detection algorithms. *arXiv* **2022**, arXiv:2205.08383. [CrossRef]
23. Mucha, W.; Kampel, M. Depth and thermal images in face detection detailed comparison between image modalities. In Proceedings of the 2022 the 5th International Conference on Machine Vision and Applications (ICMVA), Singapore, 18–20 February 2022; pp. 16–21. [CrossRef]
24. Zhang, S.; Yang, Y.; Chen, C.; Zhang, X.; Leng, Q.; Zhao, X. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and prospects. *Expert Syst. Appl.* **2023**, *237*, 121692. [CrossRef]
25. Pagé Fortin, M.; Chaib-draa, B. Multimodal multitask emotion recognition using images, texts, and tags. In Proceedings of the ACM Workshop on Crossmodal Learning and Application, Ottawa, ON, Canada, 10 June 2019; pp. 3–10. [CrossRef]
26. Aslam, A.; Sargano, A.; Habib, Z. Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks. *Appl. Soft Comput.* **2023**, *144*, 110494. [CrossRef]
27. Alsaadawi, H.; Das, R. Multimodal Emotion Recognition Using Bi-LG-GCN for the MELD Dataset. *Balk. J. Electr. Comput. Eng. (BAJECE)* **2024**, *12* .
28. Liu, Z.; Huang, G.; Chu, D.; Sun, Y. PSRMER: Proactive Services Recommendation Driven-by Multimodal Emotion Recognition. In Proceedings of the 2023 IEEE International Conference on Web Services (ICWS), Chicago, IL, USA, 2–8 July 2023; pp. 514–525. [CrossRef]
29. Mohammad, A.; Siddiqui, F.; Alam, M.; Idrees, S. Tri-model classifiers for EEG based mental task classification: Hybrid optimization assisted framework. *BMC Bioinform.* **2023**, *24*, 406. [CrossRef]
30. Tian, J.; Hu, D.; Shi, X.; He, J.; Li, X.; Gao, Y.; Toda, T.; Xu, X.; Hu, X. Semi-supervised Multimodal Emotion Recognition with Consensus Decision-making and Label Correction. In Proceedings of the1st International Workshop on Multimodal and Responsible Affective Computing, Ottawa, ON, Canada, 29 October 2023; pp. 67–73. [CrossRef]
31. Khalane, A.; Makwana, R.; Shaikh, T.; Ullah, A. Evaluating significant features in context-aware multimodal emotion recognition with XAI methods. *Expert Syst.* **2023**, e13403. [CrossRef]
32. Chen, S.; Tang, J.; Zhu, L.; Kong, W. A multi-stage dynamical fusion network for multimodal emotion recognition. *Cogn. Neurodynamics* **2023**, *17*, 671–680. [CrossRef] [PubMed]
33. Patnaik, S. Speech emotion recognition by using complex MFCC and deep sequential model. *Multimed. Tools Appl.* **2022**, *82*, 11897–11922. [CrossRef]
34. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.V.; Modi, A. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. *arXiv* **2022**, arXiv:2205.02455.
35. Cai, Y.; Li, X.; Li, J. Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors* **2023**, *23*, 2455. [CrossRef] [PubMed]
36. Bhattacharya, P.; Gupta, R.; Yang, Y. Exploring the contextual factors affecting multimodal emotion recognition in videos. *IEEE Trans. Affect. Comput.* **2021**, *14*, 1547–1557. [CrossRef]
37. Zhang, K.; Li, Y.; Wang, J.; Wang, Z.; Li, X. Feature fusion for multimodal emotion recognition based on deep canonical correlation analysis. *IEEE Signal Process. Lett.* **2021**, *28*, 1898–1902. [CrossRef]
38. Shaikh, T.; Khalane, A.; Makwana, R.; Ullah, A. Evaluating Significant Features in Context-Aware Multimodal Emotion Recognition with XAI Methods. *Authorea Preprints* **2023**. [CrossRef]
39. Zhang, X.; Li, M.; Lin, S.; Xu, H.; Xiao, G. Transformer-based Multimodal Emotional Perception for Dynamic Facial Expression Recognition in the Wild. *IEEE Trans. Circuits Syst. Video Technol.* 2023, *early access*. [CrossRef]
40. Nanduri, V.; Sagiri, C.; Manasa, S.; Sanvithatesh, R.; Ashwin, M. A Review of multi-modal speech emotion recognition and various techniques used to solve emotion recognition on speech data. In Proceedings of the 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 3–5 August 2023; pp. 577–582. [CrossRef]

41. Savci, P.; Das, B. Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages. *J. King Saud Univ.—Comput. Inf. Sci.* **2023**, *35*, 227–237. [CrossRef]
42. Liu, X.; Xu, Z.; Huang, K. Multimodal Emotion Recognition Based on Cascaded Multichannel and Hierarchical Fusion. *Comput. Intell. Neurosci.* **2023**, *2023*, 9645611. [CrossRef] [PubMed]
43. Sankala, S.; Shaik Mohammad Rafi, B.; Sri Rama Murty, K. Multi-Feature Integration for Speaker Embedding Extraction. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7957–7961. [CrossRef]
44. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv* **2016**, arXiv:1606.06259.
45. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower Provost, E.; Kim, S.; Chang, J.; Lee, S.; Narayanan, S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
46. Filali, H.; Riffi, J.; Boulealam, C.; Mahraz, M.A.; Tairi, H. Multimodal Emotional Classification Based on Meaningful Learning. *Big Data Cogn. Comput.* **2022**, *6*, 95. [CrossRef]
47. Yousaf, A.; Umer, M.; Sadiq, S.; Ullah, D.S.; Mirjalili, S.; Rupapara, V.; Nappi, M. Emotion Recognition by Textual Tweets Classification Using Voting Classifier(LR-SGD). *IEEE Access* **2020**, *9*, 6286–6295. [CrossRef]