electronics

MDPI

Article

An Emotional AI Chatbot Using an Ontology and a Novel Audiovisual Emotion Transformer for Improving Nonverbal Communication

Yun Wang [1], Liege Cheung [2], Patrick Ma [3], Herbert Lee [4] and Adela S.M. Lau [1,*]

[1] Data Science Lab, Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
[2] College of Liberal Arts & Sciences, University of Illinois Urbana-Champaign, 601 E John St, Champaign, IL 61820, USA
[3] Marvel Digital Ai Limited, 6/F, 16E, 16 Science Park East Avenue, Hong Kong Science Park, Shatin, N.T., Hong Kong
[4] Xtreme Business Enterprises Limited, 6/F, 16E, 16 Science Park East Avenue, Hong Kong Science Park, Shatin, N.T., Hong Kong
* Correspondence: adelalau@hku.hk

Abstract

One of the key limitations of AI chatbots is the lack of human-like nonverbal communication. Although there are many research studies on video or audio emotion recognition for detecting human emotions, there is no research that combines video, audio, and ontology methods to develop an AI chatbot with human-like communication. Therefore, this research aims to develop an audio-video emotion recognition model and an emotion-ontology-based chatbot engine to improve human-like communication with emotion detection. This research proposed a novel model of cluster-based audiovisual emotion recognition for improving emotion detection with both video and audio signals and compared it with existing methods using video or audio signals only. Twenty-two audio features, the Mel spectrogram, and facial action units were extracted, and the last two were fed into a cluster-based independent transformer to learn long-term temporal dependencies. Our model was validated on three public audiovisual datasets: RAVDESS, SAVEE, and RML. The results demonstrated that the accuracy scores of the clustered transformer model for RAVDESS, SAVEE, and RML were 86.46%, 92.71%, and 91.67%, respectively, outperforming the existing best model with accuracy scores of 86.3%, 75%, and 60.2%, respectively. An emotion-ontology-based chatbot engine was implemented to make inquiry responses based on the detected emotion. A case study of the HKU Campusland metaverse was used as proof of concept of the emotional AI chatbot for nonverbal communication.

Keywords: audiovisual emotion recognition; emotion ontology; AI chatbot; artificial intelligence; nonverbal communication

check for updates

Academic Editor: George A. Tsihrintzis

Received: 16 September 2025
Revised: 27 October 2025
Accepted: 27 October 2025
Published: 31 October 2025

Citation: Wang, Y.; Cheung, L.; Ma, P.; Lee, H.; Lau, A.S.M. An Emotional AI Chatbot Using an Ontology and a Novel Audiovisual Emotion Transformer for Improving Nonverbal Communication. *Electronics* **2025**, *14*, 4304. https://doi.org/10.3390/electronics14214304

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Emotion recognition has garnered significant interest from the metaverse community due to its capability of recognizing, interpreting, processing, and understanding emotions from computer images [1–5]. Human emotions can be expressed through multiple channels, including audiovisual channels such as text, speech, hand gestures, and facial expressions and biological channels such as electroencephalogram (ECG) data, body temperature, and breathing rate. These multimodal approaches have recently become popular as a way to capture meaningful feature representations for emotion detection [6–11]. However,

current methods of multimodal emotion detection have several drawbacks. First, the modality in the electroencephalogram channel and the text channel can be unobtainable and misleading, respectively. For example, the text sentence "I need to go home" can be uttered in an angry manner when you are wronged by allegations or can be expressed with excitement when you leave work before Christmas Eve. Therefore, the combination of audio and video modalities for emotion detection has been studied by many researchers [12]. Second, models that incorporate multiple modalities are usually complex and require high computational time. Many researchers have attempted to increase the number of layers in neural networks or to include modalities other than speech and visual expressions, such as ECG and thermal distributions, for higher accuracy. Nevertheless, in real applications, high model complexity reduces model practicability. Besides, current deep learning models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks suffer from low accuracy because of their inability to handle long-term dependencies.

A transformer has the capability to learn long-term temporal dependencies [13], which has proven to be appropriate for modeling emotion sequences [7]. Existing studies generally combine a standard transformer with other deep neural networks, such as CNNs, RNNs, and BiLSTM networks, for a dual-stream emotion recognition framework [14,15] or use pretrained transformer-based models for audiovisual emotion recognition [16,17], which increases the model complexity. Therefore, in this research, we aimed to develop a novel model of audio-video emotion recognition by combining the audio and video features in a clustered transformer structure in a lightweight framework. In the following sections, we review the related literature in Section 2 and present our proposed model and research methods in Section 3. We present the datasets, experiments, results, and discussion in Section 4. We present the emotion-ontology-based chatbot engine in Section 5. Finally, the conclusions and future research directions are presented in Section 6.

2. Literature Review

Producing human-like communication and interactions is the challenge in metaverse development. Communication can take place in a verbal, textual, visual, or nonvisual format. Emotion is a kind of nonverbal communication. It can be visualized through facial, vocal, or gestural behaviors. Affect units (AUs) are the fundamental building blocks of a facial action coding system (FACS), which classifies human emotions based on the facial deformation of each facial muscle [18]. Each AU represents a particular facial movement and is typically associated with a change in facial expression. The set of activated AUs can be used to construct a facial expression sequence and to determine the facial emotion [19]. Thus, AUs provide important visual information for autonomous deep learning feature extraction.

The methods for analyzing emotions can be classified into category-based and dimension-based methods. A category-based method [20] is a kind of discrete emotion analysis method, and its affects can be classified into six discrete states, namely anger, fear, sadness, enjoyment, disgust, and surprise, with each state related to a unique facial trait. In contrast, a dimension-based method is a kind of continuous emotion analysis method [21], and the circumplex model of affect is represented as a continuous valence that indicates how pleasant or unpleasant an emotion is for a person or, in an arousal space, how intense or weak an emotion is. The existing emotion datasets were based on these two analysis methods for classifying emotions into six groups and for revealing valence-arousal relationships.

To detect emotions, most researchers focused on unimodal methods, such as speech emotion recognition (SER) and facial emotion recognition (FER). Researchers used various

methods, including support vector machines (SVMs), CNNs, LSTM networks, and some novel deep learning (DL) models for SER [22–24]. Another group of researchers employed different DL models, such as CNNs and LSTM networks, to extract features from video frames for FER [25–27]. Some researchers also attempted to use a FACS to extract AUs as visual information for autonomous DL feature extraction to classify emotions [28]. However, a single modality is not enough to distinguish between complex emotion classes [29,30]. Research has shown that 55% of emotions are visual, 38% are vocal, and 7% are verbal [31]. As a result, some researchers recently focused on bimodal emotion detection models by using audio and video data.

The methods for emotion classification can be classified into machine learning and deep learning methods. Earlier studies relied on machine learning methods, such as SVMs and Gaussian mixture models (GMMs) [14,15]. However, deep learning approaches have shown superior performance because they can automatically extract translational invariant and robust features to simplify feature extraction and improve representation learning for downstream tasks [30]. Tzirakis et al. (2017) [32] employed a CNN to extract features from speech and a deep 50-layer ResNet for the visual modality and implemented an LSTM network for model-level fusion. Mocanu and Tapu (2022) [33] developed a framework that integrates spatial and temporal attention into a visual 3D-CNN and temporal attention into an audio 2D-CNN to capture intramodal features, achieving competitive results on the RAVDESS dataset. Cornejo and Pedrini (2019) [8] developed and evaluated a hybrid deep CNN to elicit discrete emotions based on audiovisual data. However, these deep learning methods have substantial computational complexity. Therefore, a lightweight and robust approach is necessary to make the training more efficient.

A transformer is a deep learning model that has recently become popular. Its model architecture relies entirely on an attention mechanism to draw global dependencies between input and output. An internal attention mechanism helps to eliminate the recurrence of any type of processing. Several existing studies highlighted the significance of a transformer architecture in emotion recognition [34,35]. Huang et al. (2020) [35] utilized a transformer model to fuse audiovisual modalities at the model level and incorporated an LSTM network to further improve continuous emotion recognition. The improved results confirmed the efficacy of transformers in extracting salient features for audiovisual emotion detection. However, all these methods combined standard transformers with other DL models, thus increasing the model complexity. In contrast, transformers use an O (1) maximum path length and have an outstanding ability to extract long-term dependencies [36]. In addition, transformer-based models can be designed in a lightweight framework to improve the effectiveness of feature extraction. Finally, apart from leveraging deep learning approaches to extract deep features, shallow features, particularly hand-crafted features in the time and frequency domains, have decisive impacts on the performance of machine learning (ML) methods [31]. Therefore, this research explores the effectiveness of purely standard transformers with a lightweight framework in recognizing video emotions by combining audio and video features.

To make communication more human-like, the detected emotion from an audio-video recognizer can be used as an emotion indicator for deducing the type of nonverbal communication. In the real world, users always interact with each other using language and sentence structures that depend on the context of the conversation and the emotion of the user. However, in the virtual world, most conversation chatbots already have predefined sentence structures and language that make the communication robotic and not human-like. Ontologies represent the semantic meaning of concepts and combine different object entities to represent concepts in communication [37–39]. They allow users to structure ill-structured communication into well-organized interactions and conduct reasoning and

deduction for communication and decision making [40]. Therefore, an ontology is a good language representation for structuring conversations based on different contexts and situations, such as human communication. An emotion ontology can be used in this study to navigate the conversation path based on the detected emotion by selecting different communication languages and sentences to make communication between a human and a chatbot more human-like. The attributes of emotions, such as happiness and sadness, and their behavioral entities, such as emotion response action, analysis, and consequence, can be used to construct an emotion ontology for conversations. In an emotion ontology, the ontology rules can be designed to reflect the logical thinking underlying communication and the reasoning from action to analysis and then to consequence based on the detected emotion [40]. Therefore, an emotion-ontology-based chatbot engine was designed in this study to create objects and rules for human-like conversations.

3. Proposed Methodology

This section discusses the research methodology of our novel cluster-based audio-video transformer for emotion recognition. The framework consists of two transformers, one for facial features and one for audio features, with a late fusion strategy, as shown in Figure 1. The preprocessing details of the audio and facial features are explained in Sections 3.1 and 3.2. The architecture of our cluster-based audio-video transformer for emotion detection is presented in Section 3.3.

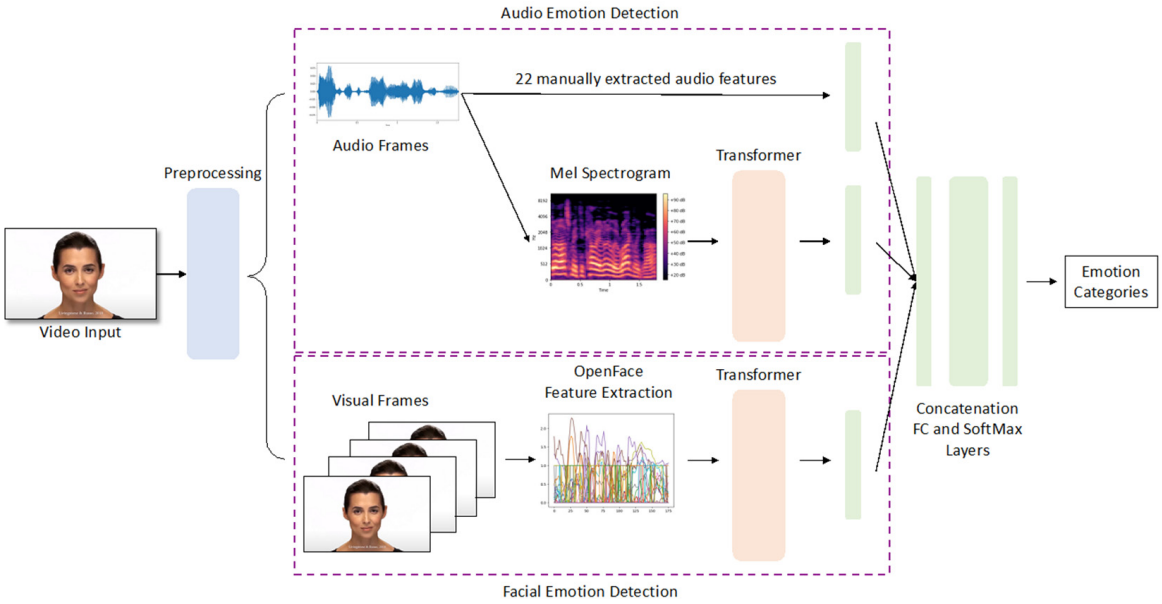


Figure 1. The architecture of our proposed transformer-based multimodal emotion detector.

3.1. Facial Feature Preprocessing

For video emotion detection, all the frames of the video clips and their 18 facial action units (AUs) were extracted using the OpenFace toolkit [34]. The OpenFace API extracted the 18 AUs and their intensity patterns [34]. The presence of the 18 AUs was binary encoded as 0 or 1, and the intensity was continuously encoded with a continuous variable between 0 to 5. Depending on the length of each video, the number of extracted frames, and hence the length of the extracted sequence of AUs, can be different. Padding techniques were employed for samples with fewer frames than their maximum (e.g., 157 frames for RAVDESS).

3.2. Audio Feature Preprocessing

For audiovisual emotion detection, the first step was to extract audio signals from the video clips for audio feature extraction. The silent parts before and after the spoken sentences were eliminated. To remove the influence of noise, speech enhancement with additive white Gaussian noise (AWGN) was applied to avoid misclassifications of positive emotions as negative emotions [17,41]. In the next step, we divided the audio signals into shorter overlapping frames with an equal length of 20 ms and 25% overlap. As the length of the audio samples differed, we applied paddings to ensure that all samples had the same length as the longest one.

Feature extraction in speech emotion detection is essential to reduce computational errors, computational time, and model complexity [42]. It is critical to extract acoustic features that can provide valid information about emotions. Therefore, as shown in Table 1, 22 features, including zero-crossing rate, chroma, and Mel-Frequency Cepstral Coefficients (MFCCs), were extracted using the Librosa library [43]. The time and frequency domains of these features reflected the reception pattern of sound frequencies intrinsic to humans, and they have been demonstrated to be robust and powerful in emotion detection. The use of several audio features, instead of just one, enables the integration of different sound characteristics, such as pitch, timbre, and harmony, into one training utterance. A richer description of audio samples tends to improve the performance of speech emotion recognition models [44]. In addition, the Mel spectrogram was extracted to highlight the low-frequency information of human auditory perception. Since the resolution at the low frequency is higher than that at the high frequency, which accords with human hearing perception, the Mel spectrogram has been widely utilized and has achieved outstanding performance in speech recognition [45]. Both the 22 features and the Mel spectrogram are further utilized as inputs for the transformers for emotion recognition.

Table 1. 22 manually extracted audio features.

Feature Name	Description
flatness	A measure to quantify how noise-like a sound is, as opposed to being tone-like.
zerocr	The rate at which a signal changes from positive to zero to negative or vice versa.
meanMagnitude	The mean of the magnitude of the spectrogram.
maxMagnitude	The maximum magnitude of the spectrogram.
stdMagnitude	The standard deviation of the magnitude of the spectrogram.
meancent	The mean of the spectral centroid, which indicates the location of the center of mass of the spectrum.
maxcent	The maximum spectral centroid.
stdcent	The standard deviation of the spectral centroid.
pitchmean	The mean of the pitch, which is the interpolated frequency estimate of a particular harmonic.
pitchmax	The maximum pitch.
pitchmin	The minimum pitch.
pitchstd	The standard deviation of the pitch.
pitch_tuning_offset	The tuning offset (in fractions of a bin) relative to 440.0 Hz.
meanrms	The mean of the root-mean-square (RMS) energy.
maxrms	The maximum root-mean-square (RMS) energy.
stdrms	The standard deviation of the root-mean-square (RMS) energy.
mfccs	The mean of MFCCs, which are derived from a cepstral representation of the audio speeches.
mfccmax	The maximum MFCC.
mfccsstd	The standard deviation of MFCCs.
chroma	Chroma represents the tonal content of a musical audio signal in a condensed form.
mel_mean	The mean of the Mel spectrogram.
contrast	Contrast is the difference between instrument sounds.

3.3. Transformer-Based Audiovisual Network Architecture

The architecture of the proposed cluster-based audio-video transformer for emotion detection is shown in Figure 1. Since our work focuses on audio and visual modalities of video inputs, we designed two independent channels (i.e., two clusters), an audio emotion detection channel (cluster) and a facial emotion detection channel (cluster), respectively. The video input was first preprocessed into images and audio signals according to the steps described in Sections 3.1 and 3.2 before sending them to the two transformers (cluster). Second, we manually selected 22 audio features for audio emotion detection. The Mel spectrogram of the audio frames was used as the other input of the transformer to extract salient audio features automatically to supplement the manually selected audio features. For the images of facial emotion detection, we applied another transformer to extract the facial features using the OpenFace toolkit. Finally, all features were concatenated and passed into a fully connected layer before yielding the eventual emotion classification result. In fact, the architecture design of our clustered model can be further extended to multiple modalities, such as ECG and textual information, for emotion recognition, and the design is not limited to audio and image emotion recognition.

3.4. System Architecture of the Emotion-Ontology-Based Chatbot Engine

To implement the emotion-ontology-based chatbot engine, OWL/RDL was used to document the emotion ontology and the concern ontology. The components of the ontologies included aspect, action, and consequence (see Figure 2). The aspect entity described the emotion attributes and the concern attributes. The action entity consisted of two entities, the response entity and the analysis entity, and was associated with the aspect entity. The response entity defined the rules and attributes of the instant responses based on the emotion or concern detected. The analysis entity defined the rules and attributes of the analysis to be performed based on the emotion or concern detected. The therapy/recommendation entity defined the rules and attributes of the therapy to be conducted based on the analysis and responses of the action entity.

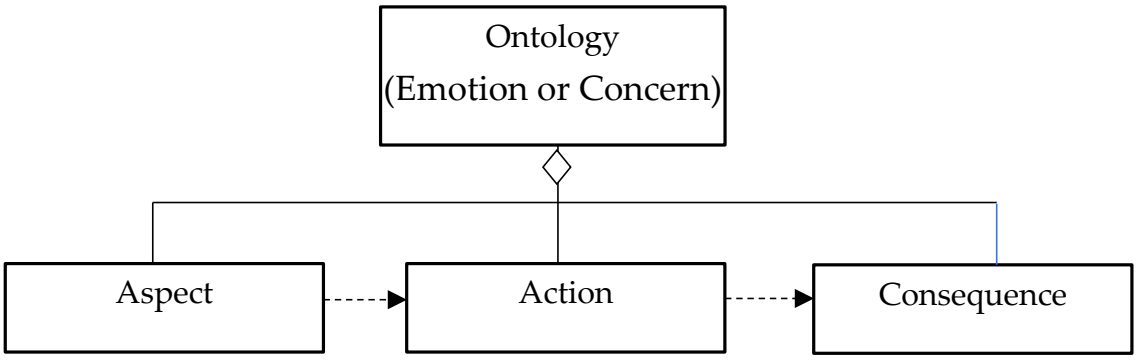


Figure 2. Ontology skeleton for the emotion and concern AI chatbot.

The emotion ontology was designed with three class entities: the emotion aspect entity, the emotion action entity, and the emotion consequence entity (see Table 2). The emotions that were proposed by Ekman [20], including anger, fear, sadness, enjoyment, disgust, surprise, and neutrality, were used to construct the attributes of the emotion aspect entity. We only changed enjoyment to be happiness in our study. The emotion aspect entity was further broken down into a positive sentiment entity, a negative sentiment entity, and a neutral sentiment entity. Second, the emotion action entity included two entities, the emotion response and the emotion analysis, and was associated with the emotion aspect entity for taking actions. The emotion response entity included attributes such as greeting

words, comfort words, and encouraging words. The emotion analysis entity included attributes such as depression, anxiety disorders, panic disorders, mood disorders, autism, dementia, psychotic disorders, and schizophrenia. Finally, the emotion consequence entity included the emotion therapy entity and was associated with the emotion action entity. The emotion therapy entity included attributes such as a slogan recommendation, a song recommendation, a video recommendation, and a doctor recommendation.

Table 2. The attributes of the emotion ontology.

Step 1: Construct the Components of the Emotion Aspect in the Emotion Ontology.	**It contains a Positive Sentiment Entity, a Negative Sentiment Entity, and a Neutral Sentiment Entity.**
The attributes of the positive sentiment entity.	Happiness, surprise.
The attributes of the neutral sentiment entity.	Neutrality.
The attributes of the negative sentiment entity.	Anger, fear, sadness, disgust.
Step 2: Construct the components of the emotion action in the emotion ontology.	**It contains the emotion response entity and the emotion analysis entity.**
The attributes of the emotion response entity.	Greeting words, comfort words, encouraging words, etc.
The attributes of the emotion analysis entity.	Depression, anxiety disorders, panic disorders, mood disorders, autism, dementia, psychotic disorders, schizophrenia, etc.
Step 3: Construct the components of the emotion consequence in the emotion ontology.	**It contains an emotion therapy entity.**
The attributes of the emotion therapy entity.	Slogan recommendation, song recommendation, video recommendation, and doctor recommendation.

To test how the detected emotion can be applied to nonverbal communication, we used the HKU Campusland metaverse as a pilot case. We developed an intelligent chatbot for program admission inquiries in the HKU Campusland. The program admission ontology contained three entities: the concern aspect entity, the action aspect entity, and the consequence aspect entity (see Table 3). First, the concern aspect entity in the program admission ontology included entities such as admission requirement inquiries, job inquiries, and program inquiries. The admission requirement entity included attributes such as DSE results, subjects to be taken, English score, Chinese score, math score, and minimal score. The job entity included attributes such as the job industry, required hard and soft skills, working experience, subjects taken, and certificates. The program inquiry entity included attributes such as program department, required hard and soft skills, job opportunities, internship opportunities, exchange opportunities, scholarship opportunities, and capstone opportunities. Second, the concern action entity included the concern response entity and the concern analysis entity and was associated with the concern aspect entity for rule deduction in the concern action entity. The concern response entity included attributes such as offices, booklets, persons, further information requirements, and appointments. The concern analysis entity included attributes such as types of jobs, types of progress, types of certificates, and types of courses. Lastly, the concern consequence ontology included a concern recommendation entity and was associated with the emotion action entity for recommendation decision making. The concern recommendation entity included program recommendation, job recommendation, admission recommendation, etc.

Table 3. The attributes of the concern ontology.

Step 1: Construct the Components of the Concern Aspect in the Program Admission Ontology.	**It contains the Admission Requirement Inquiry Entity, the Job Inquiry Entity, and the Program Inquiry Entity.**
The attributes of the admission requirement inquiry entity.	DSE results, subjects to be taken, English score, Chinese score, math score, minimal score, etc.
The attributes of the job inquiry entity.	Job industry, required hard and soft skills, working experience, subjects taken, certificates, etc.
The attributes of the program inquiry entity.	Program's department, required hard and soft skills, job opportunities, internship opportunities, exchange opportunities, scholarship opportunities, capstone opportunities, etc.
Step 2: Construct the components of the concern action in the program admission ontology.	**It contains the concern response entity and the concern analysis entity.**
The attributes of the concern response entity.	Offices, booklets, persons, further information requirements, appointments, etc.
The attributes of the concern analysis entity.	Types of jobs, types of progress, types of certificates, types of courses, etc.
Step 3: Construct the components of the concern consequence in the program admission ontology.	**It contains the concern recommendation entity.**
The attributes of the concern recommendation entity.	Program recommendation, job recommendation, admission recommendation, etc.

4. Experimental Results

4.1. Datasets

The proposed emotion recognition algorithm was evaluated on three datasets: SAVEE, RAVDESS, and RML [15,46,47]. The data preprocessing pipeline described in Sections 3.1 and 3.2 was applied to each sample of the three datasets.

- The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset [15] consists of recordings from four native male British actors at the University of Surrey in seven distinct emotions: neutrality, happiness, sadness, anger, fear, disgust, and surprise. Each actor contributes 120 English utterances, resulting in a total of 480 video samples.
- The Ryerson Audio-Visual Database of Emotion Speech and Song (RAVDESS) dataset [46] contains audio-video clips from 24 professional actors (12 females, 12 males) vocalizing two lexically matched statements in a neutral North American accent. Eight emotion classes are covered: neutrality, calmness, happiness, sadness, anger, fear, surprise, and disgust. In our study, we selected only the speech part of the dataset, resulting in 2880 video samples.
- The Ryerson Emotion Lab (RML) dataset [47] contains 720 video samples from eight actors speaking six languages (English, Mandarin, Urdu, Punjabi, Persian, and Italian). Various accents of English and Mandarin are included. Six principal emotions are delivered: happiness, sadness, anger, fear, surprise, and disgust.

4.2. Parameter Configurations

Hyperparameter tuning is an essential aspect of the deep learning process to achieve accurate predictions or classifications. Common hyperparameters include learning rate, weight decay, number of hidden neurons, and dropout rate. Grid search, the most frequently used approach for hyperparameter tuning, aims to compute the optimum values of hyperparameters exhaustively. We used RandomizedSearchCV to randomly select a prede-

fined number of hyperparameter combinations for the optimum values of hyperparameters for audio-only, video-only, and audiovisual models.

The audio-only speech recognition model performs best with 2 hidden layers and 1024 hidden neurons. The best choices of batch size and number of epochs are 32 and 500, respectively. The grid search with the Adam optimization algorithm suggests a learning rate of 0.001 and a weight decay of 1×10^{-5}. The dropout rate is tuned between 10% and 90%, with the best performance resulting from a 10% dropout rate. Similarly, we tuned the video-only and audiovisual models, whose hyperparameters are presented in Table 4.

Table 4. Optimum hyperparameters of the audio-only, video-only, and audiovisual models.

	Audio-Only	**Video-Only**	**Audiovisual**
Hidden layers	2	3	2
Hidden neurons	1024	256	128
Epochs	500	500	250
Batch size	32	32	32
Learning rate	0.001	1×10^{-5}	0.001
Weight decay	1×10^{-5}	1×10^{-5}	1×10^{-5}
Dropout rate	0.1	0.1	0

The three datasets above were used to evaluate the performance of our proposed model. Following a random stratified split mechanism, we used 80% of the samples to train the model, while the remaining 20% of the data were used for testing purposes. We used a K-fold cross validation to validate the model. For each dataset, we performed three kinds of experiments: speech emotion recognition (SER), facial emotion recognition (FER), and audiovisual emotion recognition. SER ignores the facial emotion detection channel, and FER ignores the audio emotion detection channel, as shown in Figure 1. Each experiment was conducted five times, and the average performance will be reported in Section 4.4.

4.3. Evaluation Metric

In this paper, the performance of our proposed model and baseline models was evaluated by classification accuracy. Accuracy, which is defined as the ratio of the number of accurate predictions to the total number of predictions, is computed using the equation below. The accuracy measure looks at true positives and true negatives so that it works well if the dataset is balanced.

$$Accuracy = \frac{Correct\ Predictions}{All\ Predictions}$$

4.4. Results

Since the proposed transformer-based multimodal emotion recognition framework consists of two channels for extracting audio and facial features independently, we first compared the accuracy of SER, FER, and audiovisual emotion recognition on the SAVEE, RAVDESS, and RML datasets, yielding values of 83.33%, 81.25%, and 75% and values of 90.62%, 75.69%, and 86.11%, respectively. As shown in Table 5, the average accuracy of the target emotions of audiovisual emotion recognition using our proposed model for SAVEE, RAVDESS, and RML is 92.71%, 86.46%, and 91.67%, respectively. The performance of audiovisual emotion recognition consistently outperforms SER and FER for all datasets, which is intuitive because a single modality tends to possess limited features for recognizing

emotions accurately. Also, as demonstrated in Figure 3, the diagonal entries of the bottom confusion matrix are more significant.

Table 5. SER, FER, and audiovisual classification accuracies for SAVEE, RAVDESS, and RML.

	SAVEE	RAVDESS	RML
SER	83.33%	81.25%	75%
FER	90.62%	75.69%	86.11%
Audiovisual	92.71%	86.46%	91.67%

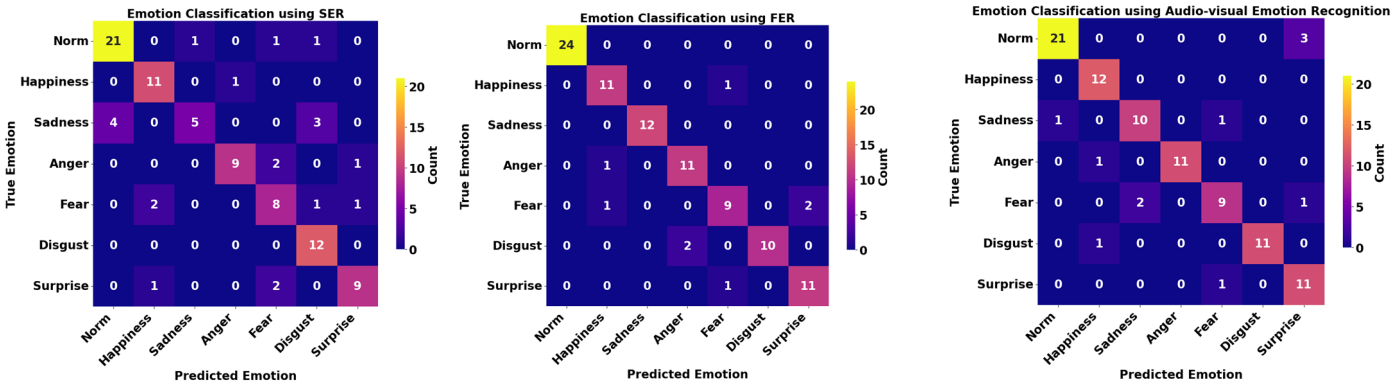


Figure 3. Confusion matrices for (**left**) SER, (**middle**) FER, and (**right**) audiovisual emotion recognition (SAVEE).

When comparing only SER with FER for all datasets, we cannot conclude whether facial or audio features are more representative in emotion detection tasks. Some emotions, such as happiness, were more accurately classified, as indicated in the top two confusion matrices in Figures 3–5. Facial muscles may express happiness better than audio signals such as tone and sound volume. Other emotions, such as disgust, have more audio variation, resulting in more subtle emotion nuances, as shown in the confusion matrices for all datasets. Therefore, incorporating both audio and facial features in emotion recognition can effectively capture all subtle changes in voice and facial landmarks so as to improve the model's performance.

In Table 6, the proposed model was compared with state-of-the-art approaches on the SAVEE, RAVDESS, and RML datasets. It is obvious that our proposed model achieved an outstanding performance for all three datasets. In Tables 7–9, the precise, recall and F1 score of each emotion detection methods were calculated with the SAVEE, RAVDESS, and RML datasets. In Table 10, it summarized the macro-precise, recall and F1 score of the 3 emotion detection methods with all datasets. In Figure 6, it shows the distribution of each of the three datasets.

For SAVEE dataset (see Table 7), the prediction performance of happiness (0.846, 0.857, 0.923), anger (0.8181, 0.880, 0.957), and disgust (0.828, 0.909, 0.957) from audio-visual method is better than that of SER and FER. The prediction performance of fear (0.640, 0.783, 0.783) is the same as that of FER, but better than that of SER. The prediction performance of sadness (0.556, 1, 0.833) and surprise (0.783, 0.880, 0.815) from audio-visual method is worse than that of FER, but better than that of SER.

For RAVDESS dataset (see Table 8), the prediction performance of happiness (0.738, 0.914, 0.963), anger (0.861, 0.737, 0.905), disgust (0.871, 0.850, 0.911), sadness (0.707, 0.586, 0.779), and fear (0.767, 0.613, 0.794) from audio-visual method is better than that of SER and FER. The prediction performance of surprise (0.795, 0.649, 0.753) from audio-visual method is worse than that of SER and better than that of FER.

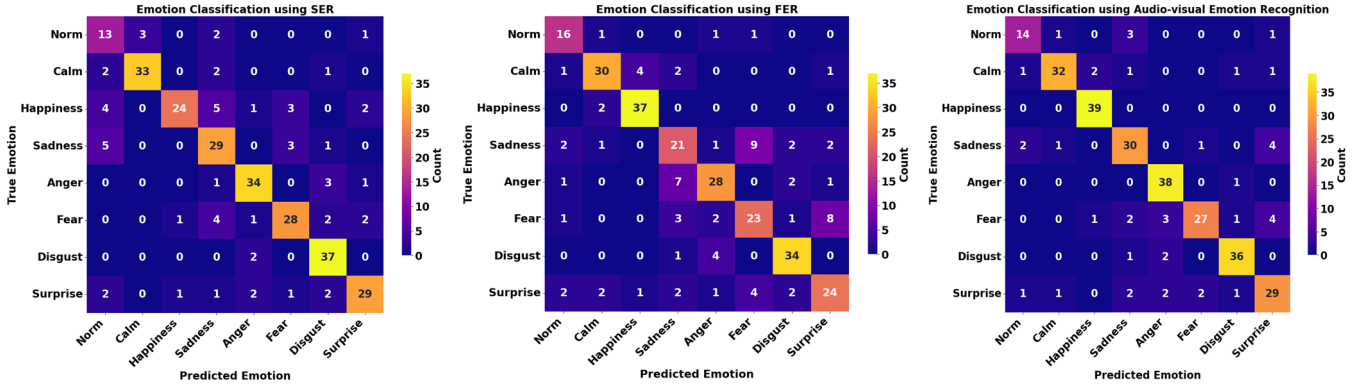


Figure 4. Confusion matrices for (**left**) SER, (**middle**) FER, and (**right**) audiovisual emotion recognition (RAVDESS).

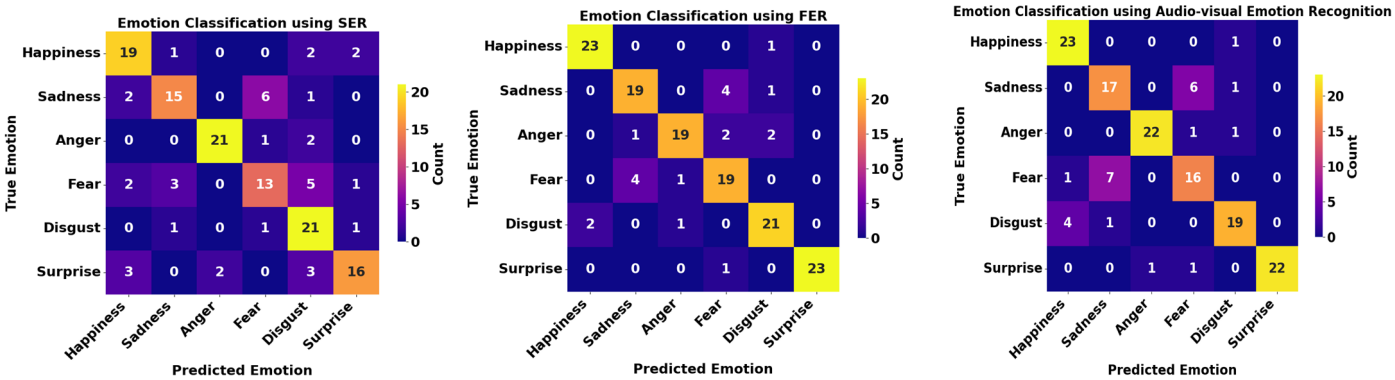


Figure 5. Confusion matrices for (**left**) SER, (**middle**) FER, and (**right**) audiovisual emotion recognition (RML).

Table 6. Comparison of various recent papers for audiovisual emotion recognition on three datasets. The highest accuracies for the three datasets are in bold.

Dataset	Literature	Year	Model	Accuracy (%)
SAVEE	Seo et al. [48]	2020	VACNN	75.00
	Singh et al. [49]	2020	RNN	77.38
	Farooq et al. [50]	2020	DCNN + CFS + SVM	82.10
	Li et al. [45]	2022	CNN + SSAE	88.96
	-	-	Our Method	**92.71**
RAVDESS	Mansouri-Benssassi et al. [51]	2019	SNN + MFCC	86.30
	Singh et al. [49]	2020	RNN	64.15
	Ghaleb et al. [52]	2020	KNN + LSTM	67.70
	Farooq et al. [50]	2020	DCNN + CFS + SVM	81.30
	Middya et al. [9]	2022	ConvLSTM2D + 1D CNN	86.00
	-	-	Our Method	**86.46**
RML	Avots et al. [53]	2018	SVM + AlexNet	60.20
	Rahdari et al. [54]	2019	NB + FRNN + SMO + BG + RF	73.12
	Li et al. [45]	2022	CNN + SSAE	83.18
	-	-	Our Method	**91.67**

Table 7. Precision, Recall and F1 for SER, FER and audio-visual emotion recognitions using SAVEE dataset.

	SAVEE-SER			SAVEE-FER			SAVEE Audio-Visual		
Class	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Happiness	0.786	0.917	0.846	0.846	0.917	0.880	0.857	1.000	0.923
Sadness	0.833	0.417	0.556	1.000	1.000	1.000	0.833	0.833	0.833
Anger	0.900	0.750	0.818	0.846	0.917	0.880	1.000	0.917	0.957
Fear	0.615	0.667	0.640	0.818	0.750	0.783	0.818	0.750	0.783
Disgust	0.706	1.000	0.828	1.000	0.833	0.909	1.000	0.917	0.957
Surprise	0.818	0.750	0.783	0.846	0.917	0.880	0.733	0.917	0.815
Norm	0.840	0.875	0.857	1.000	1.000	1.000	0.955	0.875	0.913

Table 8. Precision, Recall and F1 for SER, FER and audio-visual emotion recognitions using RAVDESS dataset.

	RAVDESS-SER			RAVDESS-FER			RAVDESS Audio-Visual		
Class	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Happiness	0.923	0.615	0.738	0.881	0.949	0.914	0.929	1.000	0.963
Sadness	0.659	0.763	0.707	0.583	0.553	0.568	0.769	0.789	0.779
Anger	0.850	0.872	0.861	0.757	0.718	0.737	0.844	0.974	0.905
Fear	0.800	0.737	0.767	0.622	0.605	0.613	0.900	0.711	0.794
Disgust	0.804	0.949	0.871	0.829	0.872	0.850	0.900	0.923	0.911
Surprise	0.829	0.763	0.795	0.667	0.632	0.649	0.744	0.763	0.753
Norm	0.500	0.684	0.578	0.696	0.842	0.762	0.778	0.737	0.757
Calm	0.917	0.868	0.892	0.833	0.789	0.811	0.914	0.842	0.877

Table 9. Precision, Recall and F1 for SER, FER and audio-visual emotion recognitions using RML dataset.

	RML-SER			RML-FER			RML Audio-Visual		
Class	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Happiness	0.731	0.792	0.760	0.920	0.958	0.939	0.821	0.958	0.885
Sadness	0.750	0.625	0.682	0.792	0.792	0.792	0.680	0.708	0.694
Anger	0.913	0.875	0.894	0.905	0.792	0.844	0.957	0.917	0.936
Fear	0.619	0.542	0.578	0.731	0.792	0.760	0.667	0.667	0.667
Disgust	0.618	0.875	0.724	0.840	0.875	0.857	0.864	0.792	0.826
Surprise	0.800	0.667	0.727	1.000	0.958	0.979	1.000	0.917	0.957

Table 10. Macro-precision, recall and F1 for SER, FER and audio-visual emotion recognitions using all datasets.

	All Datasets—SER			All Datasets—FER			All Datasets—Audio-Visual		
Emotion	**Macro-Precision**	**Macro-Recall**	**Macro-F1**	**Macro-Precision**	**Macro-Recall**	**Macro-F1**	**Macro-Precision**	**Macro-Recall**	**Macro-F1**
Happiness	0.813	0.775	0.782	0.882	0.941	0.911	0.869	0.986	0.924
Sadness	0.747	0.602	0.648	0.792	0.781	0.786	0.761	0.777	0.769
Anger	0.888	0.832	0.858	0.836	0.809	0.820	0.934	0.936	0.932
Fear	0.678	0.648	0.662	0.724	0.716	0.719	0.795	0.709	0.748
Disgust	0.709	0.941	0.807	0.890	0.860	0.872	0.921	0.877	0.898
Surprise	0.816	0.727	0.768	0.838	0.836	0.836	0.826	0.865	0.842

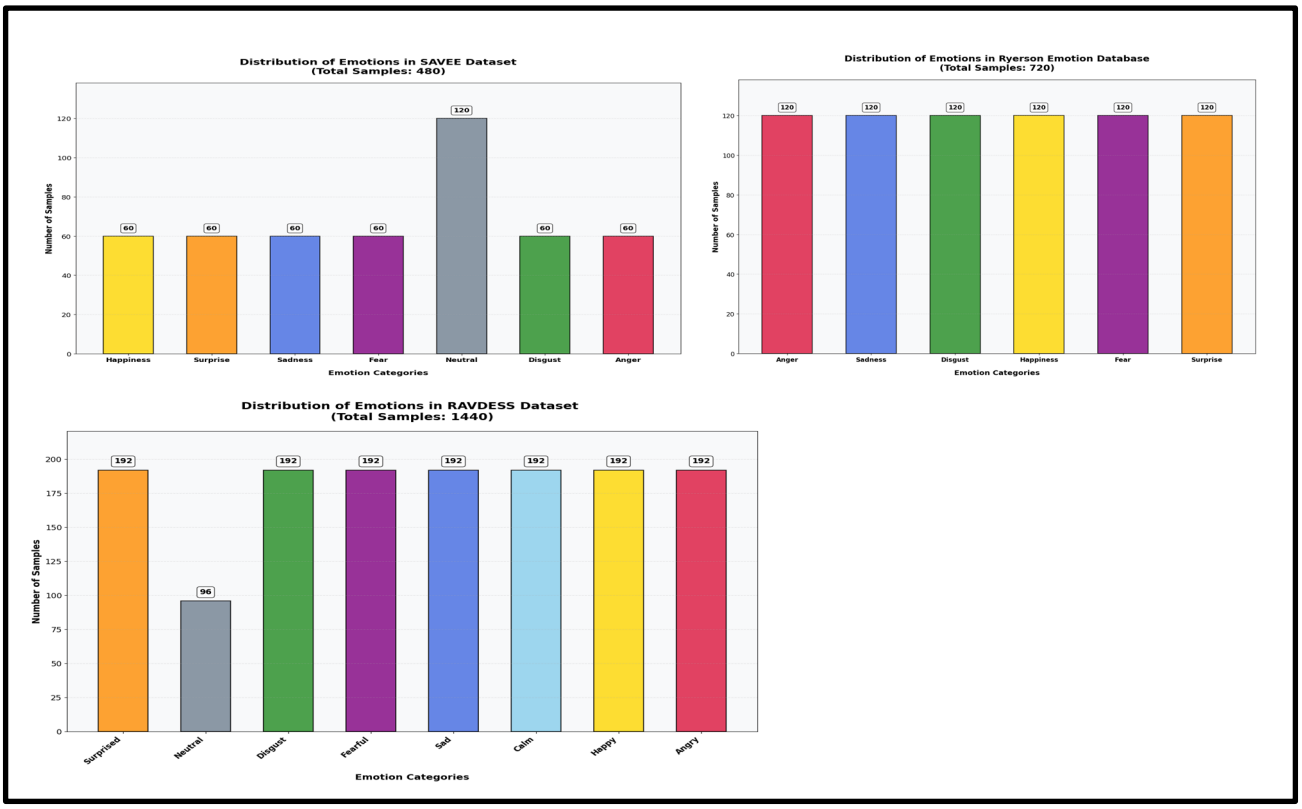


Figure 6. Data distribution of SAVEE (**left top**) and RAVDESS (**bottom**), RML (**right top**) datasets.

In RML dataset (Table 9), the prediction performance of anger (0.894, 0.844, 0.936) from audio-visual method is better than that of SER and FER. The prediction performance of happiness (0.760, 0.939, 0.885), disgust (0.724, 0.854, 0.826), sadness (0.682, 0.792, 0.694), fear (0.578, 0.760, 0667), and surprise 0.727, 0.979, (0.957) from audio-visual method is worse than that of FER, but better than that of SER.

For the three integrated datasets (see Table 10), the prediction performance of happiness (0.782, 0.911, 0.924), anger (0.858, 0.820, 0.932), fear (0.662, 0.719, 0.748), disgust (0.807, 0.872, 0.898), and surprise (0.768, 0.836, 0.842) from audio-visual method is better than that of SER and FER. The prediction performance of sadness (0.648, 0.786, 0.769) from audio-visual method is worse than that of FER and better than that of SER. Since RAVDESS dataset (1440 records) and the three integrated datasets (2640 records) have more records than the SAVEE (480 records) and RMI (720 records) datasets, the results from the RAVDESS and the three integrated datasets are more reliable and they both showed that our proposed audio-visual method outperformed than the FER and SER methods for emotion recognitions. Hence, our cluster-based transformer-based emotion recognition model has markedly improved the accuracy of video emotion detection.

5. Using an Emotion Ontology for Intelligent Responses

In order to produce a human-like emotional response in the metaverse, the next step in this research was to build an emotion ontology. The following figures show an overview of the emotion ontology skeleton (see Figure 7) and the emotion ontology schema (see Figure 8) that we designed. The implementation of the emotion ontology schema is shown in Figure 9.

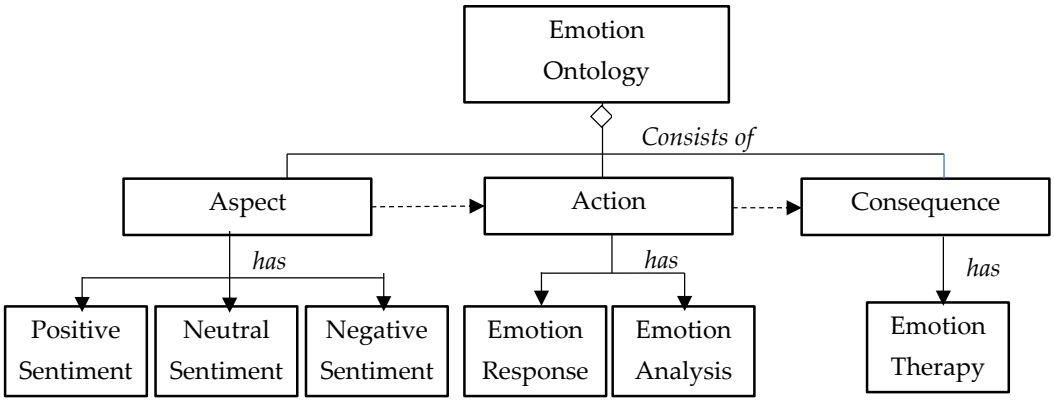


Figure 7. An overview of the emotion ontology skeleton.

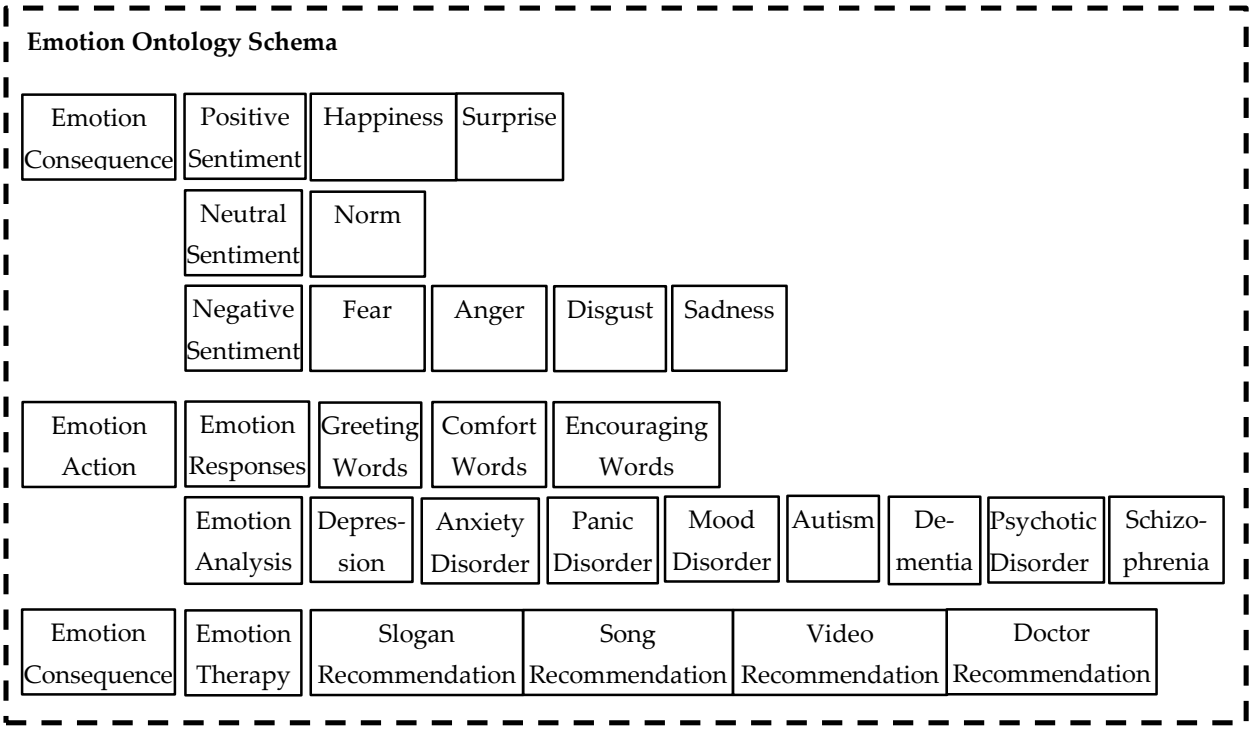


Figure 8. The emotion ontology schema.

The detected emotion is used to determine the emotion actions to be taken, such as saying encouraging words to the candidate if a sad sentiment is detected. Further analysis of the participant's emotion is performed in the emotion analysis entity and then passed to the emotion therapy entity for taking a recommendation decision, such as saying "Never give up."

HKU Campusland is a metaverse for students, faculties, and companies to interact and communicate with nonverbal communication. It allows participants to interact with our chatbot to make admission inquiries, program inquiries, and job inquiries. To make the AI chatbot more human-like, the output of the audio-video emotion recognition apps will be fed into our AI chatbot for more human-like communication. The AI chatbot was constructed with a program admission ontology that was used as a backbone for future development. The following figures show an overview of the program admission ontology skeleton (see Figure 10) and the program admission ontology schema (see Figure 11) that we designed.

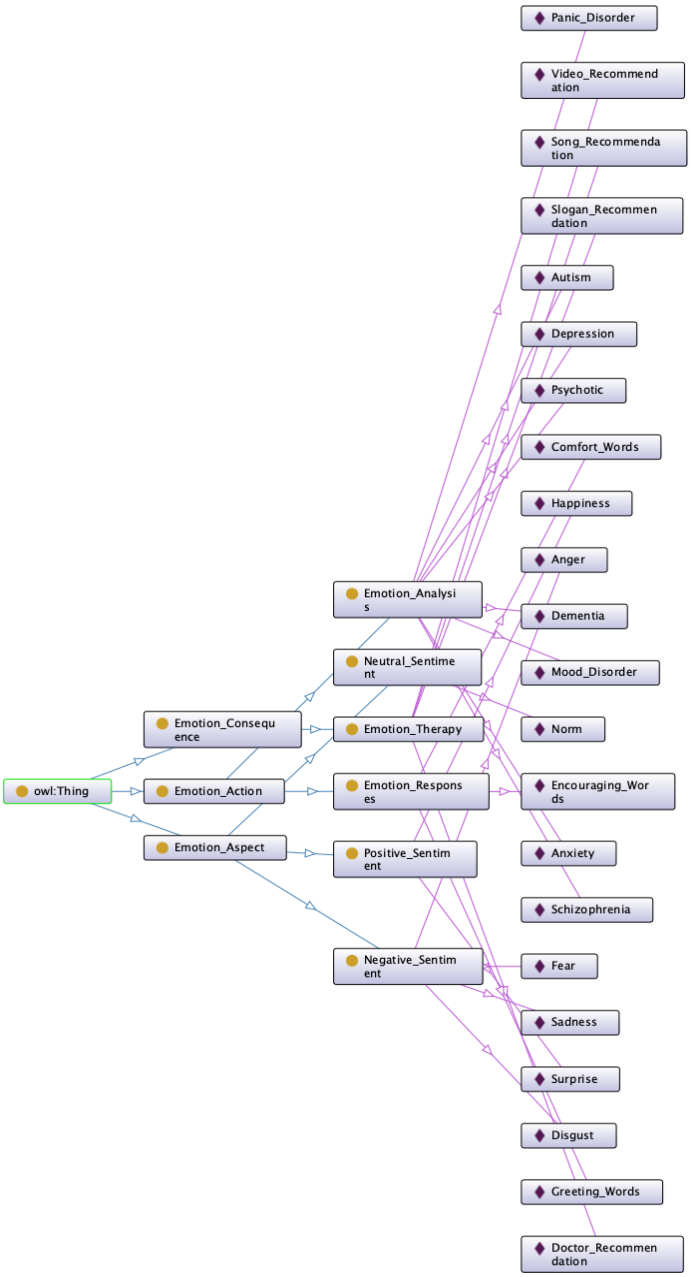


Figure 9. The implementation of the emotion ontology schema.

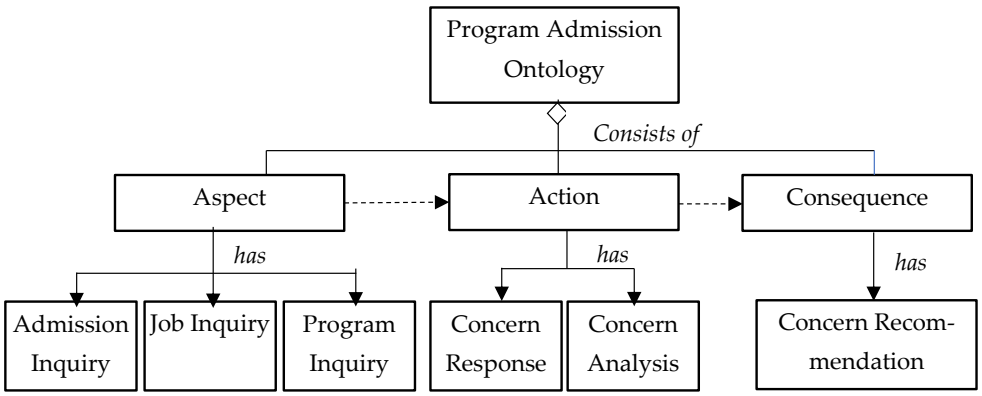


Figure 10. An overview of the program admission ontology skeleton.

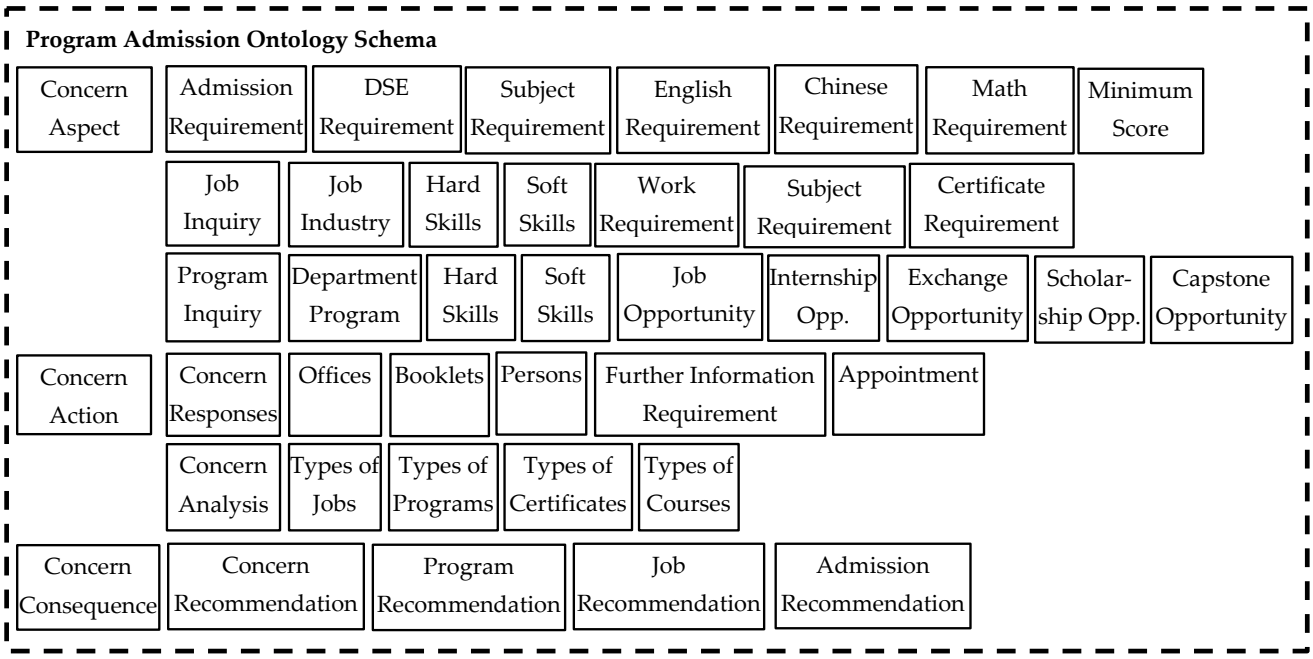


Figure 11. The program admission ontology schema.

The schema of the program admission ontology in Figure 11 was designed as a knowledge base for answering candidates' program inquiries and job inquiries. The implementation of the program admission ontology schema is shown in Figure 12. The candidate might ask questions about program admission, program details, and graduate jobs. The AI chatbot will first give an emotional response to comfort the candidate and will then answer the candidate's questions. To answer a question, the chatbot will first classify the questions into admission, job, or program questions. It will next determine the corresponding responses and perform further analysis of the types of jobs, programs, certificates, or courses that are suitable for the candidate, and it will finally offer recommendations regarding programs, jobs, or admission. Since the AI methods for questioning and answering are not the main topic of this research, the details on how to build the recommender are not discussed here. As a proof of concept, we only illustrate how the detected emotion can be further used in nonverbal communication in the metaverse.

In Figure 13, we provide an example scenario to demonstrate how we use the detected emotion for program inquiries. In this example, using the audio-video emotion recognition app, we detected that the candidate was sad in the emotion aspect module and established that the candidate suffered from an anxiety disorder in the emotion analysis module. We therefore first recommended the slogan "never give up" in the emotion consequence module. In the concern response, we determined that the question belonged to program inquiries in the concept aspect module. We selected "professor for consultation" in the concern action module and established that the inquiry was about what program to study. We then recommended the data science and statistics programs to the candidate in the concern consequence module. Since this research is about a novel model of audio-video emotion recognition and how it can be used in a chatbot with human-like emotions and nonverbal communication, the AI models and rule-based engine of the AI chatbot that govern the responses to questions are not presented and discussed here.

The example of using SPARQL, Jena, and OWL/RDF/XML for reasoning responses to user enquiry of the above example is shown in Figures 14–18.

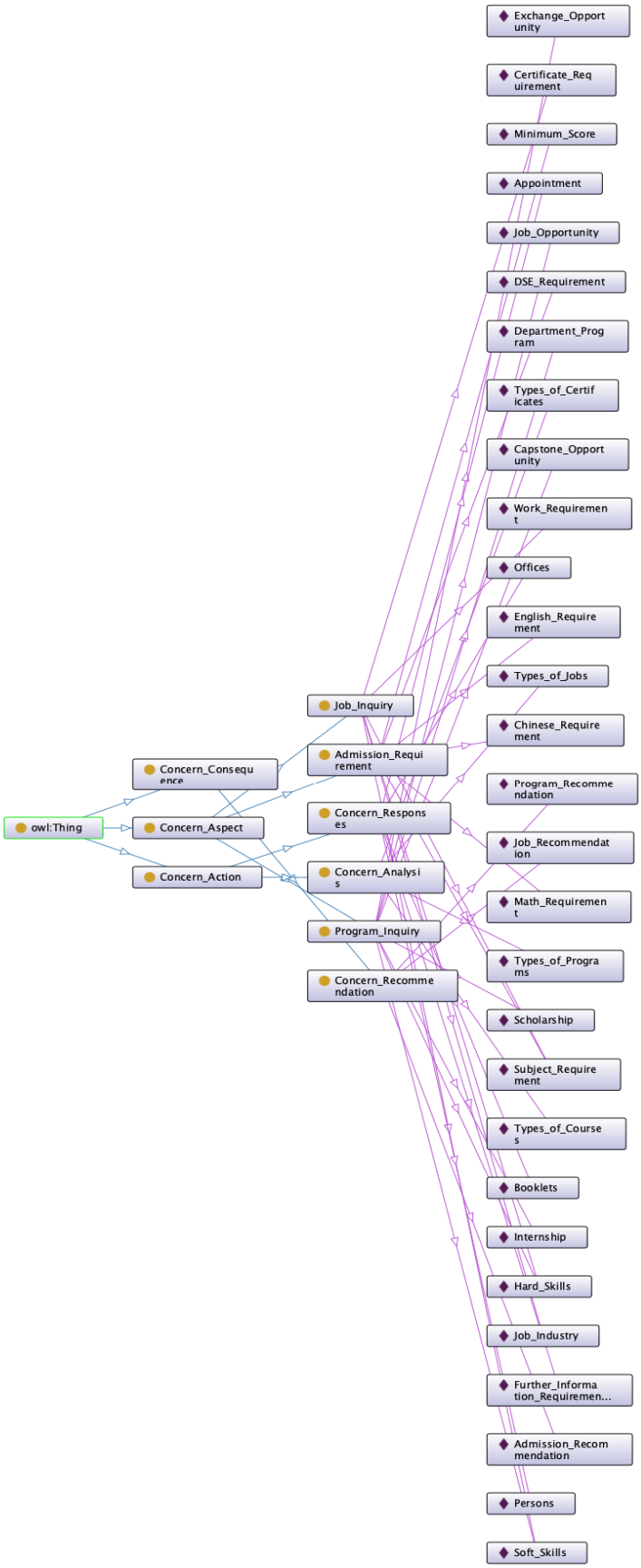


Figure 12. The implementation of the program admission ontology schema.

Example Scenario:

Bob (in a sad voice): I got a marginal score in English at the DSE result. But I got a good grade in math. May I ask which program I can study in your department with the highest chance of being admitted? I am so worried about whether your school can admit me.

NPC (the ontology engine): Never give up! You can contact Dr Lau at the HKU SAAS Data Science Lab. There are some data science and statistics programs in the Department of Statistics and Actuarial Science that may be suitable to you.

Figure 13. An example scenario for an AI chatbot with emotion analysis.

Reasoning pseudocode:

1. Define the Fuseki SPARQL endpoint (loaded with emotion_ontology and program_admission_ontology; configured in local environment)
2. Define SPARQL query to implement reasoning, for example:
    a. Link user's emotion (Sadness → Negative_Sentiment) to Emotion_Therapy (Slogan_Recommendation)
    b. Link user's disorder (Anxiety → Emotion_Analysis) to Slogan_Recommendation (infer slogan "Never give up!")
    c. Link user's math score (good → Math_Requirement → Admission_Requirement) to Program_Inquiry
    d. Infer Program_Recommendation from Program_Inquiry (Data Science and Statistics programs)
    e. Infer contact person (Dr Lau) from Program_Recommendation
3. Execute SPARQL query via Python's SPARQLWrapper
4. Parse JSON results and generate the NPC response as per the scenario

Figure 14. Reasoning pseudocode for the example of AI chatbot emotion analysis.

Reasoning SPARQL code:

```
reasoning_query = f"""
{self.prefixes}
SELECT ?slogan ?program ?contact
WHERE {{
# Reasoning 1: Negative emotion + Emotion Analysis → Slogan Recommendation
emotion:{user_emotion} rdf:type emotion:Negative_Sentiment .
emotion:{user_disorder} rdf:type emotion:Emotion_Analysis .
emotion:Slogan_Recommendation rdf:type emotion:Emotion_Therapy .
BIND("Never give up!" AS ?slogan) # Inferred from Slogan_Recommendation
# Reasoning 2: Good math score → Match Math Requirement → Program Recommendation
program:Math_Requirement rdf:type program:Admission_Requirement .
program:Program_Recommendation rdf:type program:Concern_Recommendation .
program:Program_Inquiry rdf:type program:Concern_Aspect . # User's question type
BIND("Data Science and Statistics programs" AS ?program) # Inferred from Math_Requirement
BIND("Dr. Lau at HKU SAAS Data Science Lab" AS ?contact) # Linked to Program_Recommendation
}} LIMIT 1
"""
```

Figure 15. Reasoning SPARQL code for the example of AI chatbot emotion analysis.

Jena server configuration:

```
@prefix :              <http://base/#> .
@prefix fuseki:    <http://jena.apache.org/fuseki#> .
@prefix rdf:         <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix tdb2:       <http://jena.apache.org/2016/tdb#> .    # TDB2 prefix (not TDB1)

# Define the service with TDB2 support
:myService rdf:type fuseki:Service ;
      fuseki:name "emotion_program" ;    # Your dataset name
      fuseki:serviceQuery "query" ;       # Enable query
      fuseki:serviceUpdate "update" ;    # Enable updates
      fuseki:serviceUpload "upload" ;    # Enable uploads
      fuseki:serviceReadWriteGraphStore "data" ;    # Enable Graph Store Protocol
      fuseki:dataset :myDataset .

# Define TDB2 dataset (instead of TDB1)
:myDataset rdf:type tdb2:DatasetTDB2 ;    # Use TDB2 type
      tdb2:location "databases/emotion_program" .    # Same storage location
```

Figure 16. Jena server configuration for the AI chatbot emotion analysis.

OWL code:

"emotion_ontology.owl":

```
<?xml version="1.0"?>
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
    ……..
   <Declaration>
      <Class IRI="#Emotion_Action"/>
   </Declaration>
    ……
   <Declaration>
      <NamedIndividual IRI="#Video_Recommendation"/>
   </Declaration>
   <SubClassOf>
      <Class IRI="#Emotion_Analysis"/>
      <Class IRI="#Emotion_Action"/>
   ……
   <SubClassOf>
      <Class IRI="#Positive_Sentiment"/>
      <Class IRI="#Emotion_Aspect"/>
   </SubClassOf>
   <DisjointClasses>
      <Class IRI="#Emotion_Action"/>
      <Class IRI="#Emotion_Aspect"/>
      <Class IRI="#Emotion_Consequence"/>
   </DisjointClasses>
       ……
   <DisjointClasses>
      <Class IRI="#Negative_Sentiment"/>
      <Class IRI="#Neutral_Sentiment"/>
      <Class IRI="#Positive_Sentiment"/>
   </DisjointClasses>
   <ClassAssertion>
      <Class IRI="#Negative_Sentiment"/>
      <NamedIndividual IRI="#Anger"/>
   </ClassAssertion>
     ……
   <ClassAssertion>
      <Class IRI="#Emotion_Therapy"/>
      <NamedIndividual IRI="#Video_Recommendation"/>
   </ClassAssertion>
</Ontology>
<!-- Generated by the OWL API (version 4.5.29.2024-05-13T12:11:03Z) https://github.com/owlcs/owlapi -->
```

Figure 17. Sample code of OWL/RDF/XML for emotion ontology.

"program_admission_ontology.owl":

```
<?xml version="1.0"?>
<Ontology xmlns="http://www.w3.org/2002/07/owl#"
    ……
    ……

    <Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#"/>
    <Declaration>
        <Class IRI="#Admission_Requirement"/>
    </Declaration>
    ……
    <Declaration>
        <Class IRI="#Program_Inquiry"/>
    </Declaration>
    <Declaration>
        <ObjectProperty IRI="#consistOf"/>
    </Declaration>
    <Declaration>
        <NamedIndividual IRI="#Admission_Recommendation"/>
    </Declaration>
    ……
    <Declaration>
        <NamedIndividual IRI="#Work_Requirement"/>
    </Declaration>
    <SubClassOf>
        <Class IRI="#Admission_Requirement"/>
        <Class IRI="#Concern_Aspect"/>
    </SubClassOf>
    ……
    <SubClassOf>
        <Class IRI="#Program_Inquiry"/>
        <Class IRI="#Concern_Aspect"/>
    </SubClassOf>
    <DisjointClasses>
        <Class IRI="#Admission_Requirement"/>
        <Class IRI="#Job_Inquiry"/>
        <Class IRI="#Program_Inquiry"/>
    </DisjointClasses>
    ……
    <DisjointClasses>
        <Class IRI="#Concern_Analysis"/>
        <Class IRI="#Concern_Responses"/>
    </DisjointClasses>
    <ClassAssertion>
        <Class IRI="#Concern_Recommendation"/>
        <NamedIndividual IRI="#Admission_Recommendation"/>
    </ClassAssertion>
    ……
    <ClassAssertion>
        <Class IRI="#Job_Inquiry"/>
        <NamedIndividual IRI="#Work_Requirement"/>
    </ClassAssertion>
</Ontology>
<!-- Generated by the OWL API (version 4.5.29.2024-05-13T12:11:03Z) https://github.com/owlcs/owlapi -->
```

Figure 18. Sample code of OWL/RDF/XML for program admission ontology.

6. Conclusions and Future Work

Emotion detection plays an essential role in human nonverbal communication. This paper proposed a novel clustered transformer-based lightweight framework for audiovisual emotion recognition that overcomes the challenges of long-term temporal dependencies and model complexity of traditional models. The proposed model consists of two channels for audio feature and facial feature extraction and uses a late fusion strategy to improve the model's accuracy. The proposed model was compared with state-of-the-art approaches on three public audiovisual datasets, SAVEE, RAVDESS, and RML, and achieved superior accuracy scores of 92.71%, 86.46%, and 91.67%, respectively, demonstrating that the purely transformer-based model has significantly improved the accuracy of video emotion detection. Second, the emotion-ontology-based AI chatbot engine demonstrated how the detected emotion can be used to make intelligent and emotional responses to participants in a more human-like manner of communication. In future work, to further validate our model's effectiveness, we can explore the performance on multiple modalities, including ECG and textual information. Instead of manually preprocessing audio and video input for all datasets, we can further investigate an automated way of feature extraction and analysis so that our framework can be extended to emotion detection in other applications, such as a movie recommender and the monitoring of patients' emotions in a hospital setting.

Other than that, since emotion detection requires to capture and record individuals' facial expressions or voice patterns for training or using the model, it may infringe on personal privacy, data access and data usage issues [55,56]. Therefore, a framework including data minimization, adaptive consent mechanisms, and transparent model logic is required to protect privacy and fairness use of the data [57]. A clear disclosure of the data collection purposes and its usage to users is required. A mechanism to bookkeep users' consent and providing real time feedback from users on the use of their data are required. A guideline to guide developers on matching user principles and regulatory requirements is also needed. A risk contingency plan can be developed for handling the disaster of data leakage, stolen, or unauthorized access accidently. Last but not least, using the data masking and feature extraction technologies [58,59] for implementing the emotion detection model is another method to minimize the data use, and to protect user privacy information such as their face and voice that can further research. Lastly, encrypting the data [60] for data storage and transmission can ensure data confidential and privacy.

Author Contributions: Y.W.: designed the transformer model and was responsible for data preprocessing, programming audio-visual emotion detection, and the summary of the results during her study in department of statistics and actuarial science of the University of Hong Kong. L.C.: designed and implemented the emotion ontology for the intelligent chatbot in the metaverse during his study in University of Illinios, Urban Camphaion. A.S.M.L.: initiated the project, part of the programming, supervised the first two authors, offered advice, draft and polished the paper. H.L.: validated the findings and review the paper. P.M.: initiated the topic, validated the findings, and review the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research paper writing received no external funding. The APC was funded from both Department of Statistics and Actuarial Science and HKU SAAS Data Science Lab.

Data Availability Statement: No data provided in this study for public.

Acknowledgments: We would like to acknowledge HKU SAAS Data Science Lab for providing the research collaboration environment to the authors.

Conflicts of Interest: Author Patrick Ma was employed by the company Marvel Digital Ai Limited, and author Herbert Lee was employed by the company Xtreme Business Enterprises Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Franzoni, V.; Milani, A.; Nardi, D.; Vallverdú, J. Emotional machines: The next revolution. *Web Intell.* **2019**, *17*, 1–7. [CrossRef]
2. Ayadi, M.E.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
3. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Manag.* **2001**, *18*, 32–80. [CrossRef]
4. Cornejo, J.; Pedrini, H. Bimodal Emotion Recognition Based on Audio and Facial Parts Using Deep Convolutional Neural Networks. In Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 111–117. [CrossRef]
5. Morningstar, M.; Nelson, E.E.; Dirks, M.A. Maturation of vocal emotion recognition: Insights from the developmental and neuroimaging literature. *Neurosci. Biobehav. Rev.* **2018**, *90*, 221–230. [CrossRef]
6. Luna-Jiménez, C.; Griol, D.; Callejas, Z.; Kleinlein, R.; Montero, J.M.; Fernández-Martínez, F. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. *Sensors* **2021**, *21*, 7665. [CrossRef] [PubMed]
7. Siriwardhana, S.; Kaluarachchi, T.; Billinghurst, M.; Nanayakkara, S. Multimodal Emotion Recognition with Transformer-Based Self Supervised Feature Fusion. *IEEE Access* **2020**, *8*, 176274–176285. [CrossRef]
8. Cornejo, J.Y.R.; Pedrini, H. Audio-Visual Emotion Recognition Using a Hybrid Deep Convolutional Neural Network based on Census Transform. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 3396–3402. [CrossRef]
9. Middya, A.I.; Nag, B.; Roy, S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. *Knowl. Based Syst.* **2022**, *244*, 108580. [CrossRef]
10. Hao, M.; Cao, W.-H.; Liu, Z.-T.; Wu, M.; Xiao, P. Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. *Neurocomputing* **2020**, *391*, 42–51. [CrossRef]
11. Oh, J.M.; Kim, J.K.; Kim, J.Y. Multi-Detection-Based Speech Emotion Recognition Using Autoencoder in Mobility Service Environment. *Electronics* **2025**, *14*, 1915. [CrossRef]
12. Ramyasree, K.; Kumar, C.S. Expression Recognition Survey Through Multi-Modal Data Analytics. *International. J. Comput. Sci. Netw. Secur.* **2022**, *22*, 600–610. Available online: http://paper.ijcsns.org/07_book/202206/20220674.pdf (accessed on 26 October 2025).
13. Andayan, F.; Lau, B.T.; Tsun, M.T.; Chua, C. Hybrid LSTM-Transformer Model for Emotion Recognition from Speech Audio Files. *IEEE Access* **2022**, *10*, 36018–36027. [CrossRef]
14. Zhalehpour, S.; Onder, O.; Akhtar, Z.; Erdem, C.E. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Trans. Affect. Comput.* **2016**, *8*, 300–313. [CrossRef]
15. Wang, Y.; Guan, L. Recognizing Human Emotional State From Audiovisual Signals. *IEEE Trans. Multimed.* **2008**, *10*, 936–946. [CrossRef]
16. Tran, M.; Soleymani, M. A Pre-Trained Audio-Visual Transformer for Emotion Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; pp. 4698–4702. [CrossRef]
17. Lee, S.; Yu, Y.; Kim, G.; Breuel, T.; Kautz, J.; Song, Y. Parameter Efficient Multimodal Transformers for Video Representation Learning. *arXiv* **2020**. [CrossRef]
18. Baltrusaitis, T.; Mahmoud, M.; Robinson, P. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; pp. 1–6. [CrossRef]
19. Bagheri, E.; Esteban, P.G.; Cao, H.-L.; De Beir, A.; Lefeber, D.; Vanderborght, B. An Autonomous Cognitive Empathy Model Responsive to Users' Facial Emotion Expressions. *ACM Trans. Interact. Intell. Syst.* **2020**, *10*, 1–23. [CrossRef]
20. Ekman, P. Basic Emotions. In *Handbook of Cognition and Emotion*; John Wiley & Sons, Ltd.: Chichester, UK, 2005; pp. 45–60. [CrossRef]
21. Posner, J.; Russell, J.A.; Peterson, B.S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **2005**, *17*, 715–734. [CrossRef]
22. Šumak, B.; Brdnik, S.; Pušnik, M. Sensors and Artificial Intelligence Methods and Algorithms for Human–Computer Intelligent Interaction: A Systematic Mapping Study. *Sensors* **2022**, *22*, 20. [CrossRef] [PubMed]
23. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2018**, *20*, 1576–1590. [CrossRef]
24. Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schuller, B. Speech Emotion Classification Using Attention-Based LSTM. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1675–1685. [CrossRef]
25. Gupta, A.; Arunachalam, S.; Balakrishnan, R. Deep self-attention network for facial emotion recognition. *Procedia Comput. Sci.* **2020**, *171*, 1527–1534. [CrossRef]

26. Yu, Z.; Liu, G.; Liu, Q.; Deng, J. Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing* **2018**, *317*, 50–57. [CrossRef]
27. Norhikmah; Lutfhi, A.; Rumini. The Effect of Layer Batch Normalization and Droupout of CNN model Performance on Facial Expression Classification. *JOIV Int. J. Inform. Vis.* **2022**, *6*, 481–488. [CrossRef]
28. Alkawaz, M.H.; Mohamad, D.; Basori, A.H.; Saba, T. Blend Shape Interpolation and FACS for Realistic Avatar. *3D Res.* **2015**, *6*, 6. [CrossRef]
29. Aouani, H.; Ayed, Y.B. Speech emotion recognition with deep learning. *Procedia Comput. Sci.* **2020**, *176*, 251–260. [CrossRef]
30. Zhang, J.; Yin, Z.; Chen, P.; Nichele, S. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf. Fusion* **2020**, *59*, 103–126. [CrossRef]
31. Mehrabian, A. Communication Without Words. In *Communication Theory*, 2nd ed.; Routledge Press: Oxfordshire, UK, 2017; Chapter 2; pp. 193–200. [CrossRef]
32. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.W.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1301–1309. [CrossRef]
33. Mocanu, B.; Tapu, R. Audio-Video Fusion with Double Attention for Multimodal Emotion Recognition. In Proceedings of the IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Nafplio, Greece, 26–29 June 2022; pp. 1–5. [CrossRef]
34. Baltrusaitis, T.; Zadeh, A.; Lim, Y.C.; Morency, L.P. OpenFace 2.0: Facial Behavior Analysis Toolkit. In Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition, Xi'an, China, 15–19 May 2018; pp. 59–66. [CrossRef]
35. Huang, J.; Tao, J.; Liu, B.; Lian, Z.; Niu, M. Multimodal Transformer Fusion for Continuous Emotion Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3507–3511. [CrossRef]
36. Khorram, S.; Aldeneh, Z.; Dimitriadis, D.; McInnis, M.; Provost, E.M. Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition. *arXiv* **2017**. [CrossRef]
37. Lau, A. Implementation of an onto-wiki toolkit using web services to improve the efficiency and effectiveness of medical ontology co-authoring and analysis. *Inform. Health Soc. Care* **2009**, *34*, 73–80. [CrossRef]
38. Tsui, E.; Wang, W.M.; Cheung, C.F.; Lau, A. A Concept-Relationship Acquisition and Inference Approach for Hierarchical Taxonomy Construction from Tags. *Inf. Process. Manag.* **2010**, *46*, 44–57. [CrossRef]
39. Lau, A.; Tse, S. Development of the Ontology Using a Problem-Driven Approach: In the Context of Traditional Chinese medicine Diagnosis. *Int. J. Knowl. Eng. Data Min.* **2010**, *1*, 37–49. [CrossRef]
40. Lau, A.; Tsui, E.; Lee, W.B. An Ontology-based Similarity Measurement for Problem-based Case Reasoning. *Expert Syst. Appl.* **2009**, *43*, 6547–6579. [CrossRef]
41. Huang, C.; Chen, G.; Yu, H.; Bao, Y.; Zhao, L. Speech emotion recognition under white noise. *Arch. Acoust.* **2013**, *38*, 457–463. [CrossRef]
42. Al-onazi, B.B.; Nauman, M.A.; Jahangir, R.; Malik, M.M.; Alkhammash, E.H.; Elshewey, A.M. Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion. *Appl. Sci.* **2022**, *12*, 9188. [CrossRef]
43. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.W.; McVicar, M.; Battenbergk, E.; Nieto, O. Librosa: Audio and Music Signal Analysis in Python. 2015. Available online: https://www.youtube.com/watch?v=MhOdbtPhbLU (accessed on 26 October 2025).
44. Wang CRen, Y.; Zhang, N.; Cui, F.; Luo, S. Speech Emotion Recognition Based on Multi-feature and Multi-lingual Fusion. *Multimed. Tools Appl.* **2022**, *81*, 4897–4907. [CrossRef]
45. Li, J.; Zhang, X.; Huang, L.; Li, F.; Duan, S.; Sun, Y. Speech Emotion Recognition Using a Dual-Channel Complementary Spectrogram and the CNN-SSAE Neutral Network. *Appl. Sci.* **2022**, *12*, 9518. [CrossRef]
46. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
47. Haq, S.; Jackson, P.J.B. Speaker-Dependent Audio-Visual Emotion Recognition. In Proceedings of the International Conference on Audio-Visual Speech Processing, Norwich, UK, 10–13 September 2009; pp. 53–58. Available online: https://www.researchgate.net/publication/358647534_Speaker-Dependent_Audio-Visual_Emotion_Recognition (accessed on 1 March 2023).
48. Seo, M.; Kim, M. Fusing Visual Attention CNN and Bag of Visual Words for Cross-Corpus Speech Emotion Recognition. *Sensors* **2020**, *20*, 5559. [CrossRef]
49. Singh, R.; Puri, H.; Aggarwal, N.; Gupta, V. An Efficient Language-Independent Acoustic Emotion Classification System. *Arab. J. Sci. Eng.* **2020**, *45*, 3111–3121. [CrossRef]
50. Farooq, M.; Hussain, F.; Baloch, N.K.; Raja, F.R.; Yu, H.; Zikria, B.Y. Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network. *Sensors* **2020**, *20*, 6008. [CrossRef]

51. Mansouri-Benssassi, E.; Ye, J. Speech Emotion Recognition With Early Visual Cross-modal Enhancement Using Spiking Neural Networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8. [CrossRef]
52. Ghaleb, E.; Popa, M.; Asteriadis, S. Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK, 3–6 September 2019; pp. 552–558. [CrossRef]
53. Avots, E.; Sapiński, T.; Bachmann, M.; Kamińska, D. Audiovisual emotion recognition in wild. *Mach. Vis. Appl.* **2019**, *30*, 975–985. [CrossRef]
54. Rahdari, F.; Rashedi, E.; Eftekhari, M. A Multimodal Emotion Recognition System Using Facial Landmark Analysis. *Iran. J. Sci. Technol. Trans. Electr. Eng.* **2019**, *43*, 171–189.
55. Achilleos, G.; Limniotis, K.; Kolokotroni, N. Exploring Personal Data Processing in Video Conferencing Apps. *Electronics* **2023**, *12*, 1247. [CrossRef]
56. Ferreira, S.; Marinheiro, C.; Mateus, C.; Rodrigues, P.P.; Rodrigues, M.A.; Rocha, N. Overcoming Challenges in Video-Based Health Monitoring: Real-World Implementation, Ethics, and Data Considerations. *Sensors* **2025**, *25*, 1357. [CrossRef] [PubMed]
57. Barker, D.; Tippireddy, M.K.R.; Farhan, A. Ethical Considerations in Emotion Recognition Research. *Psychol. Int.* **2025**, *7*, 43. [CrossRef]
58. Fitwi, A.; Chen Yu Zhu, S.; Blasch, E.; Chen, G. Privacy-Preserving Surveillance as an Edge Service Based on Lightweight Video Protection Schemes Using Face De-Identification and Window Masking. *Electronics* **2021**, *10*, 236. [CrossRef]
59. Zhang, Z.; Cilloni, T.; Walter, C.; Fleming, C. Multi-Scale, Class-Generic, Privacy-Preserving Video. *Electronics* **2021**, *10*, 1172. [CrossRef]
60. An, D.; Hao, D.; Lu, J.; Zhang, S.; Zhang, J. A Visually Meaningful Color-Video Encryption Scheme That Combines Frame Channel Fusion and a Chaotic System. *Electronics* **2024**, *13*, 2376. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.