Multimedia Tools and Applications (2022) 81:35173–35194
https://doi.org/10.1007/s11042-022-13363-4

1174: FUTURISTIC TRENDS AND INNOVATIONS IN MULTIMEDIA SYSTEMS USING BIG DATA, IOT AND CLOUD TECHNOLOGIES (FTIMS)

A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling

Sadil Chamishka[1] · Ishara Madhavi[1] · Rashmika Nawaratne[2] · Damminda Alahakoon[1] · Daswin De Silva[2] · Naveen Chilamkurti[3] · Vishaka Nanayakkara[1]

Received: 4 October 2020 / Revised: 9 March 2022 / Accepted: 3 June 2022 /
Published online: 22 June 2022

Abstract
The advancements of the Internet of Things (IoT) and voice-based multimedia applications have resulted in the generation of big data consisting of patterns, trends and associations capturing and representing many features of human behaviour. The latent representations of many aspects and the basis of human behaviour is naturally embedded within the expression of emotions found in human speech. This signifies the importance of mining audio data collected from human conversations for extracting human emotion. Ability to capture and represent human emotions will be an important feature in next-generation artificial intelligence, with the expectation of closer interaction with humans. Although the textual representations of human conversations have shown promising results for the extraction of emotions, the acoustic feature-based emotion detection from audio still lags behind in terms of accuracy. This paper proposes a novel approach for feature extraction consisting of Bag-of-Audio-Words (BoAW) based feature embeddings for conversational audio data. A Recurrent Neural Network (RNN) based state-of-the-art emotion detection model is proposed that captures the conversation-context and individual party states when making real-time categorical emotion predictions. The performance of the proposed approach and the model is evaluated using two benchmark datasets along with an empirical evaluation on real-time prediction capability. The proposed approach reported 60.87% weighted accuracy and 60.97% unweighted accuracy for six basic emotions for IEMOCAP dataset, significantly outperforming current state-of-the-art models.

Keywords Bag-of-audio-words · Machine learning · Artificial intelligence · Big data · Emotion analysis

✉ Naveen Chilamkurti
n.chilamkurti@latrobe.edu.au

Extended author information available on the last page of the article

1 Introduction

Real-time multimedia applications and services including video conferencing, telepresence, real-time content delivery, telemedicine, voice-controls on wearables and online-gaming, have contributed to the exponential growth of the Internet multimedia traffic [7]. Multimedia systems are rich sources of integrated audio, text and video streams which facilitate capturing, processing and transmission of multimedia information. This rapidly growing internet traffic of human conversations contains a massive volume of information, especially the voice-related attributes that help characterize human behaviour and embedded emotions. Emotions impact the way individuals think and act in real-life situations. Humans have unique ways to express themselves, sometimes even combining multiple emotions together as a mix [21]. These basic and complex emotion-swings influence the human physical movements, perceptions, cognition, actions and personality [22]. The ability to detect, capture and utilize human emotions from digital footprints left in multimedia has become a very important research direction.

Identification of emotions has a great value in multiple aspects. It allows us to understand the people we communicate with as decisions people make differ based on their emotions [22]. Although a precise and a concrete interpretation of how emotions are provoked in human minds is not yet available, scientists and psychologists have concentrated effort in defining and interpreting emotion generation in different perspectives including cognitive sciences, neurology, psychology and social sciences [22]. On one hand, the emotion generation can be viewed as a joint function of a physiologically arousing condition and the way a person tends to evaluate or appraise the situation. In terms of neurology, the emotions are regarded as activations caused by the changes in the density of neural stimulations or firings per unit time. In addition, emotions have been categorized as positive and negative emotions. However, the sufficiency of this segregation is questionable as the valence tag or the positivity or negativity of emotions depend on the situation encountered, and in most situations require deeper and more granular interpretation.

Recent studies have recognized the role of significant low-level acoustic features including spectrograms, Mel-Frequency Cepstral Coefficients (MFCC), fundamental frequency (F0) analysed via high-level statistical functions and other deep learnt features for emotion detection [29, 40]. In addition, deep neural networks and sequence modelling techniques are developed and evaluated for emotion detection from audio. The Recurrent Neural Network models have been widely used, due to its ability to model sequential information while preserving more complex interactions. The LSTM and GRU based neural architectures are followed by additional attention layers to precisely detect the emotional content embedded in human speech. As an embedding mechanism, the Bag-of-Audio-Words (BoAW) feature embeddings are known to perform well for detecting dimensional emotions (arousal and valence) in literature, yet its robustness is not experimented for detecting more granular categorical emotions i.e., happy, sad, angry, etc. [38]. Latest research includes the development of model architectures which are capable of utilizing the context of conversations to enhance the emotion prediction strength.

Despite the significant amount of research conducted focusing on emotion detection from audio conversations, a number of key issues are still to be addressed. The following Table 1 summarises the key limitations in emotion detection from audio conversations.

Table 1 Limitations of emotion detection from audio conversations

Limitation	Description
Limited types of emotions	Existing systems are capable of predicting only a limited number of emotions. Mixed emotion prediction has not yet been experimented and evaluated thoroughly [21].
Low performance	The relatively low performance reported for audio features compared to textual features in emotion detection tasks. Although the text and acoustic modalities contain complementary cues for emotion detection [40], the low accuracies in emotion capture from acoustic features has not yet been clearly explicated [32].
Speaker diarization	The conversations have to be split into segments based on the speaker identity to feed into the emotion prediction models.

In this context, our research presents four major contributions to the state-of-the-art models in constructing a novel feature based approach for detecting emotion categories from audio conversations.

1. First, we propose a Natural Language Processing (NLP) inspired Bag of Audio Words (BoAW) approach to represent rich audio embeddings in distinguishing six basic emotion categories, i.e., happy, sad, neutral, angry, frustrated and excited.
2. Second, alongside the BoAW feature embeddings, we propose an appropriate attention mechanism that best aligns with the feature representations input to the emotion detection model.
3. Third, we explore and evaluate the robustness and the effectiveness of this feature extraction and embedding process followed by an emotion detection model which utilizes conversation context information for emotion class predictions.
4. Fourth, we evaluate the performance of each component in the proposed approach in terms of delivering emotion predictions in real-time to validate the usefulness of the novel approach to be integrated with real-time applications or systems. This validation demonstrates the benefit of capturing emotion variations of the participants in a conversation such as in human or machine-driven (automated) call-centers and health care systems.

The rest of the paper is organized as follows. The second section briefly discusses the conceptual background of the theories adapted. Section three presents the proposed approach followed by section four on experiments and results. The last section concludes the paper, discussing the implications of the results, limitations, and potential for future work.

2 Conceptual background

Emotion detection research has developed with collaborative research contributions from psychology, cognitive science, machine learning and Natural Language Processing (NLP) [34]. Emotional intelligence (or an artificial counterpart) is important for machines when interacting with humans, as emotions and the ability to sense emotion play an important role in maintaining productive social interactions [35]. A significant volume of research has been conducted in the area of detection of emotionally relevant information from different sources, offering positive as well as compelling evidence of impact on multiple fields covering healthcare, human resource management and Artificial Intelligence [1–4, 6]. The traditional

process of emotion prediction includes frame-based feature extraction (low-level descriptors - LLDs), followed by utterance-level information collection, and input to a classification or a regression technique [36]. This includes and highlights the existing feature extraction techniques and the robustness of various models used for emotion predictions. Existing research attempts focus on audio feature extraction techniques via handcrafted methods, and deep-learnt features together with emotion detection models experimented in multimodality (text transcripts, audio, visual) aspects, utterance level or attentive (contextual) emotion predictors based on classifiers of Support Vector Machines (SVM), variations of deep learning networks, etc. Nevertheless, significant potential for further improvements exist in the accuracy and the interpretability of the proposed approaches due to the current limited capability in predicting higher variation of basic emotions [40, 41], lack of conversational context utilization [40] and inferior performance reported in acoustic-based emotion detection compared to text [32].

Prevailing feature extraction techniques either extract shallow handcrafted features and apply statistical functions including mean, variance, range, quartiles, linear regression coefficients, etc. to determine the temporal variations of feature patterns or allow deep neural networks to unveil useful feature representations. The work elaborated in [29] utilizes deep learnt features for emotion detection. The related studies attempt to find the suitability of pitch information in determining the emotions from audio segments. As Mel-scale spectrograms cause loss of pitch information, the research work in [27, 36] better utilizes pitch information by extracting linearly spaced spectrogram features. It has been reported that the usage of statistical learning of the layers in a deep neural network for feature extraction, yields better results in contrast to the handcrafted low-level features [36]. Inspired by the domain of Natural Language Processing, Bag-of-Audio-Words (BoAW) feature representations have been successfully used for classifying audio events and detecting other acoustic activities [38]. The work in [38] utilizes Mel-Frequency Cepstral Coefficients (MFCC) as low-level features for the codebook creation using a random sampling of audio words. The feature embeddings are input to a Support Vector Regression model to retrieve predictions for emotions in the dimensions of valence and arousal which respectively refer to the positive or negative affectivity and the degree of excitement of the emotion. Regardless of the rich interpretability of this feature representation, the evaluations have been conducted only in the dimensional aspects (arousal and valence) of emotions.

Most of the existing research studies focus on constructing emotion detection models. Due to the lack of publicly available datasets, most of the researchers have conducted their study on the rich Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset using sequence modelling techniques for audio and text analysis. The datasets including the Multimodal EmotionLines Dataset (MELD) in which an average of five participants were engaged in conversations have been more challenging to evaluate when compared to the dyadic conversations found in the IEMOCAP dataset. There are different groupings of existing models, including multimodal emotion detection models, classifiers involving deep neural networks and context-independent or dependent predictors. A considerable number of these attempts utilize multimodal information to construct robust solutions. In literature, the treatment of multimodal features, i.e., transcripts, audio and visual modalities, provide the potentiality to construct rich feature representations induced by their complementary information. The research conducted in [40] studies the impact of speech and text transcriptions from a speech for emotion detection through multiple CNN based architectures. This work expresses the fitness of Mel-Frequency Cepstral Coefficients (MFCC) over spectrogram features in terms of acoustic modality and the low representational capability of text embeddings due to the

possible loss of information in the speech-to-text translation process. Furthermore, a considerable accuracy gain has been attained in this research from combining text and audio modalities. A state-of-the-art multimodal emotion detection technique is proposed in [20] in which the textual features are extracted from pre-trained word embeddings via a single layer Convolutional Neural Network (CNN), the audio features being extracted from openSMILE [16] toolkit and visual modality from a deep 3D-CNN architecture. Although the multimodal emotion detection performs better than the unimodal approaches, a considerable accuracy gap prevails between the text modality and the audio modality based feature representations [32].

The current main emotion detection models are based on distance or tree-based machine learning algorithms as well as variants of deep neural networks. The research work in [25] implements a hierarchical decision tree classifier which utilizes prior knowledge (i.e., acoustic features can differentiate between high and low activated emotions) for the initial, top-level split and performing a series of classifier-results (Gaussian Mixture Models, Linear Discriminant Analysis, Support Vector Machine) for the cascading splits outperforming Support Vector Machine (SVM) baseline accuracies. Several variations of deep neural networks such as recurrent neural networks and convolutional networks have been reported [36]. The research described in [18] has proposed a Deep Neural Network (DNN) based segment level predictor on emotion probability distribution which is input to a succeeding Extreme Learning Machine (ELM), a single-hidden-layer neural network, for utterance level emotion detection. This DNN and ELM stacked approach is efficient and outperforms Support Vector Machine (SVM) based emotion detection for small scale training partitions. Not only the sequence modelling techniques but also the Convolutional Neural Network (CNN) based techniques [40] have been proposed using a combined feature flow of Mel-Frequency Cepstral Coefficients (MFCC) and spectrograms. The work in [24] proposes a novel approach to combine CNN based feature extraction from sequential data and feed to an RNN.

While most of the studies conducted focus on acquiring emotionally relevant information independently from utterances, RNN based memory networks with attention mechanisms have been proposed and successfully utilized to capture historical aspects of the conversation and query the memory bank to get relevant information needed to detect emotions [23, 33]. Due to the recurrent structure of RNN models that handle input data, RNN models have been the first choice among researchers recently, for sequence modelling tasks such as speech recognition and emotion prediction when compared to conventional Hidden Markov Models (HMM). Furthermore, the improved variations of RNN including LSTM and GRU cells have proven to be capable of pertaining short and long term context information over a conversation [26]. Thus, the ability inherent in RNN models with memory cells to track memory states in sequential data, provide valid reasons to incorporate RNN models in emotion recognition of multi-party conversations to better utilize contextual information in conversations. As human nature is naturally influenced by various emotions which arise from the context of a conversation, it is essential for a model to exhibit the capability to detect complex emotions in considerable accuracies, with better utilization of the context information in conversations. Poria et al. [14] use contextual information from neighbouring utterances of the same speaker to predict emotions. Recent emergence in the research community for filtering emotionally salient information from utterances and infusing conversational context has gained significant accuracy gains. The research work in [29] focuses on performing emotion detection by combining a bidirectional LSTM with a weighted pooling strategy using an attention mechanism which enables the network to focus on emotionally salient parts of a sentence. The

approach in [20] utilizes an RNN-based memory network, including multi-hop attention mechanism which injects self and interpersonal influences into the global memory of the conversation to acquire affective summaries of the context. The DialogueRNN and its variants (BiDialogRNN, BiDialogRNN with attention) are recently constructed state-of-the-art models used to retrieve the emotion predictions from conversations considering the context information considering the global context of the conversation, speaker states, and emotion states using three separate Gated Recurrent Units (GRU) [28]. Although the performance of the model has been investigated in terms of textual modality and trimodal scenario (text, audio, visual), no work has been reported on the individual performance of the audio modality.

Major limitations identified in the study of existing work include the low accuracy of audio modality based feature representations compared to the text modality and lack of consideration towards the capability of Natural Language Processing (NLP) influenced audio embedding techniques (BoAW) for predicting emotion categories. Furthermore, the literature review reveals the existing possibilities of utilizing contextual information in a conversation flow via RNN models and appropriate attention mechanisms to yield enriched emotion predictions. Addressing the aforementioned limitations, we design and evaluate a new audio feature extraction approach consisting of the BoAW approach and a state-of-the-art recurrent neural network to uplift the performance of emotion detection from human conversations.

3 Proposed method

The proposed methodology is empowered by the Bag of Audio Words (BoAW) based novel feature extraction approach and a state-of-the-art Recurrent Neural Network model. The method aims to highlight the applicability of the Bag of Words (BoW) feature representation techniques extensively used in the NLP domain, to represent audio features and evaluate the performance improvement of emotion classification in audio conversations utilizing one of the state-of-the-art emotion detection models. Adhering to dependencies present in human conversations, the adopted RNN model captures the long term and short term contextual information in the conversation. In a typical conversation between two or more parties, each individual implicitly contributes to the context of the conversation and the emotions of each speaker vary over time. Therefore, the feature encodings from the BoAW are combined with the contextual information derived along the conversation using the RNN, to improve the predictive power of the model. Figure 1 illustrates the proposed approach consisting of the proposed audio feature representation mechanism, the BoAW approach, from which the feature embeddings are input to the RNN model. The low-level descriptors (LLDs) related to emotional classification tasks are extracted from audio streams using the openSMILE toolkit. As the number of extracted LLD feature vectors for different utterances can vary in lengths depending on the utterance lengths, an encoding mechanism is required to represent the extracted LLDs prior to present as input to an RNN model at each step of the conversation. The BoAW approach outputs fixed-size audio embeddings for each utterance with the support of a codebook generated from the LLD features from the training partition. The codebook consists of frequently occurring distinct feature vectors from the audio segments, leading to a compact and rich feature encoding while reducing the impact of infrequent noisy audio segments. At the final stage, a Bidirectional Recurrent Neural Network model with attention is selected to retrieve the respective emotion predictions from the utterances of the conversation.

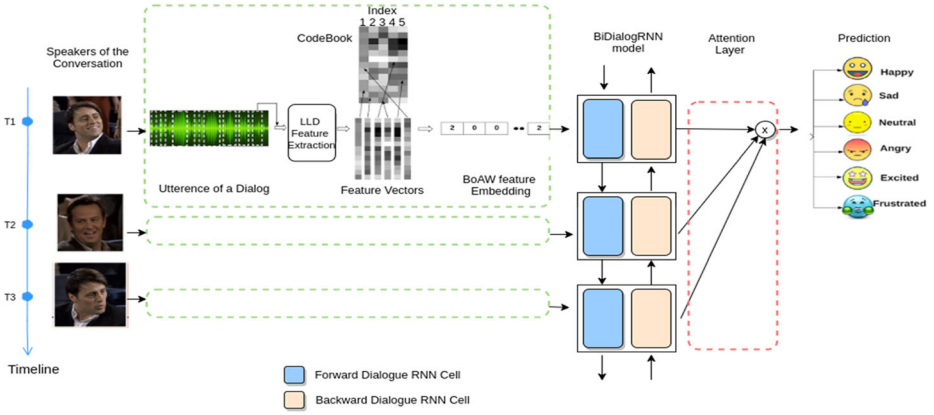

Fig. 1 Bag of Audio Words based Recurrent Neural Network model for emotion detection from audio

The model functions in three phases. (1) Feature extraction, (2) Bag of Audio words (BoAW) feature embeddings, and (3) Emotion extraction. The feature extraction phase yields 130 low-level features per every 10 milliseconds of an utterance. The BoAW component creates rich feature embeddings from the extracted low-level features, by generating a term frequency matrix which denotes the frequency of each index of the codebook within the extracted low-level features from audio data. These rich feature embeddings contribute in reducing the existing accuracy gap between text and audio feature representations. The emotion extraction is achieved through a recurrent neural network architecture and utilizes the rich feature embeddings to predict 6 basic emotion categories. We describe each of the phases in the following subsections.

3.1 Feature extraction

Extracting prominent feature sets for emotion detection is important for constructing a precise emotion detection model. As a pre-processing step, the audio conversations have to be segmented into utterances and the respective speakers should be identified. Manual speaker diarization or Automatic speaker diarization techniques can be utilized to segment and label each utterance of the speakers. Then the audio files of utterances are formatted in the form of 16-bit PCM WAV files and use the available, open-sourced openSMILE toolkit [16] to extract acoustic low-level descriptors (LLDs). The LLDs include Energy related features such as root mean square energy, zero-crossing rate and spectral features including MFCC 1–14, spectral flux along with voicing related features like fundamental frequency, logHNR, jitter, shimmer as provided in the ComParE_2016 feature set [39] which are known to be promising features which carry emotional contents in human voice. The extracted feature vector comprises 65 low-level audio descriptors and their respective first order derivatives. Altogether, 130 low level features are extracted from each 25 milliseconds with 10 milliseconds frame rate, assuming the emotion related features are stationary during the aforementioned interval. After the feature extraction phase, a rich feature corpus consisting of time-varying feature sequences is available for the downstream feature engineering tasks. In the next step, the variable-length time series feature vectors consisting of low-level features of the audio signal are put through a BoAW based rich encoding mechanism prior to feeding into prediction models.

3.2 Bag of Audio Words (BoAW) feature embeddings

The approach has its roots in NLP where the documents are represented as Bag-of-Words representations, and this approach provides Bag-of-Audio-Words as an encoding mechanism for audio data. The overall BoAW feature embedding approach is illustrated in Fig. 2. The first stage creates an indexed codebook consisting of different feature patterns from the training audio data which were generated from a set of low-level features extracted from audio. The algorithm can achieve this by randomly selecting patterns or by kmeans + + clustering algorithm [8]. The codebook vectors are iteratively selected randomly with higher dissimilarity within codebook vectors based on the Euclidean distance measure. At the presence of an unseen (test) low level feature pattern, the best matching codebook pattern, having the least distance to the pattern at hand, is selected and the term frequency of the respective index is increased. Once all the term frequencies are found in utterance levels (bag/histogram of audio words), a term frequency matrix is created, of which the entries act as fixed size encodings that can be input to a sequence to sequence modelling approach for predicting the latent emotion categories. As in the standard NLP BoW approach in document classification, the decimal logarithm is taken to shrink term frequency range as shown in Eq. (1) where TF and w denote term-frequency and audio-word respectively. The complete BoAW framework, namely the openXBOW has been implemented in Java and available in open source [37].

$$TF_{new}(w) = lg(TF(w) + 1) \tag{1}$$

3.3 Recurrent neural network model

BiDialogueRNN [28] is a recently developed model used to retrieve the emotion predictions from a conversation utilizing the context information by taking into account both the

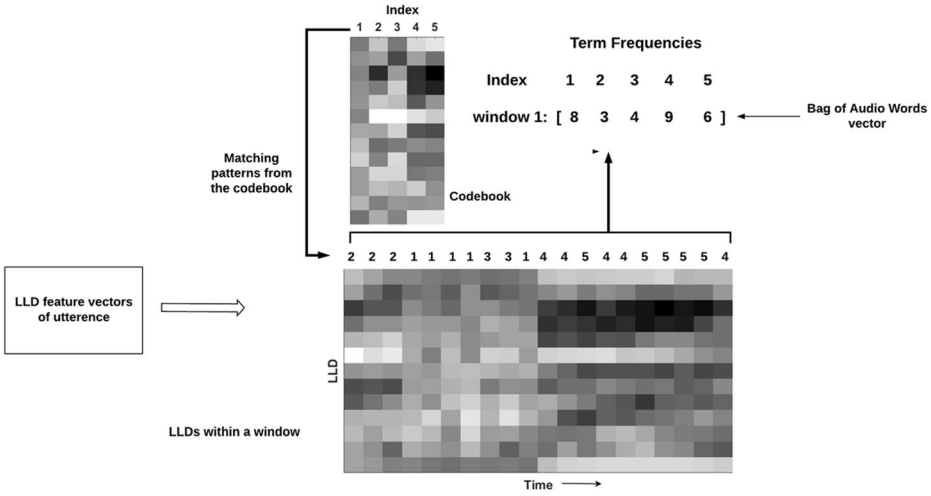

Fig. 2 Bag of Audio Words based feature embedding approach

past and the future utterances. Three major factors that help understand the emotion variations in conversations are elaborated in [28], in particular, the global context of the conversation (G_t), speaker's state (P_t) and emotion of the utterance (E_t) which are modeled using three separate recurrent neural networks. According to Chung et al. [12], GRU cells have been used to capture the long-term dependencies of sequences which leads to identifying the dynamics of the conversation while preserving the inter-party relationships by maintaining the context of the conversation. Each GRU cell computes a hidden state defined as h_t= GRU(h_{t-1}, x_t), where x_t is the current input and h_{t-1} is the previous GRU state. The state h_t also serves as the current GRU output.

3.3.1 Global state GRU

When the speakers are engaged in a conversation by taking turns, at each turn of the speaker, the context of the conversation has to be updated. Further, the most recent-past context of the conversation has a relatively high impact on the emotional state of the speaker. Therefore, it is crucial to keep track of the information of the previous states of the conversation which is achieved by encoding the utterance and speaker's previous state which are concatenated and fed to the global state GRU at each time step. The captured inter-speaker and inter-utterance dependencies convey more reliable contextual representations of the conversation as the current utterance u_t updates the previous state $q_{s(u_t),t-1}$ of the speaker to $q_{s(u_t),t}$. This information is captured through the GRU cell as shown in Eq. (2) along with g_{t-1} which is the previous context of the conversation to generate the current context of the conversation $g_t \in R^{\varrho}$ where ϱ is the size of the global state vector and $\oplus$ is the concatenation. This is illustrated in Fig. 3.

$$g_t = GRU_{global}(g_{t-1}, (u_{t-1} \oplus q_{s(u_t),t-1})) \tag{2}$$

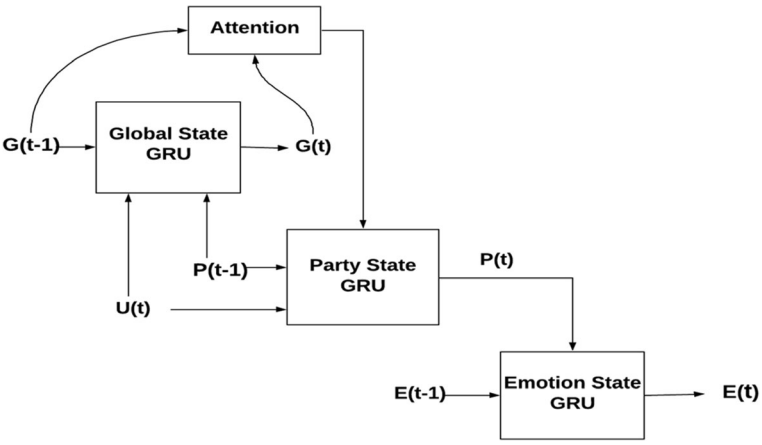

Fig. 3 BiDialogRNN with the improved attention mechanism

3.3.2 Party state GRU

The states of the individual participants change over time during the conversation and informative qualities embedded in participant states can be used to detect the emotions of the speaker. Party state GRU has proposed a computational model to capture the speaker's state throughout the conversation by updating corresponding current states of the speakers at the presence of an utterance, as shown in Eq. (3). Generally, the response of a speaker depends on the previous global states of the conversation. To cope with this feature, a speaker's previous state $q_{s(u_t),t-1}$ is modified to $q_{s(u_t),t}$ based on the current utterance u_t of the speaker and the context c_t as shown in Eq. (3). The attention mechanism proposed for the model, as an extension when utilizing BoAW approach has improved the performance, which has the ability to come up with more reliable party states form RNNs by considering the effects of hidden states. The emotionally relevant states of the conversations get high attention scores and provide contextually enriched representation c_t. In Eqs. (4) and (6), the $g_1, g_2 \ldots g_t$ are the preceding states ($g_t \in R^{\varrho}$) of the conversation, α denotes the attention score over the previous global states and $W\alpha \in R^{1\times\varrho}$ denotes a weight vector for the softmax layer.

$$q_{s(u_t),t} = GRU_{party}(q_{s(u_t),t-1}, (u_t \oplus c_t)) \tag{3}$$

$$\alpha = softmax(g_t{}^T W_\alpha [g_1, g_2 \ldots g_t]) \tag{4}$$

$$softmax(x) = [\frac{e^{x_1}}{_i e^{xi}} + \frac{e^{x_2}}{_i e^{xi}}, \ldots] \tag{5}$$

$$c_t = \alpha[g_1, g_2 \ldots g_{t-1}] \tag{6}$$

3.3.3 Proposed attention mechanism

As the speaker is influenced by previous states of the conversation, it is important to pay attention to the emotionally relevant segments of the conversation to determine the speaker's next state. We propose an attention mechanism as the most appropriate attention when coping with BoAW based feature representations. The proposed attention mechanism considers the transposed vector of the current global state g_t as shown in Eq. (4), in contrast to obtaining the transpose of the current utterance u_t as previously in [28]. Results indicate this proposed attention to be more relevant to relate to the attention weights based on g_t, as g_t contains compact (dense) information than the high dimensional (2000-length) sparse utterance encodings resulted from encoded u_t.

3.3.4 Emotion state GRU

In order to predict the emotional state e_t at timestamp t, the emotionally relevant features embedded in the party states $q_{s(u_t),t}$ are input to the emotion state GRU along with the previous

emotion state $e_{t-1}$of the speaker as indicated in Eq. (7). It can be identified as speaker GRU and global GRU in combination act similar to an encoder, whereas emotion GRU serves as a decoder. The forward and backward passes of the Bidirectional Emotion state RNN provide emotional representations of the speakers throughout the conversation. The forward and backward emotion states are concatenated, and a separate attention-based mechanism is applied to capture emotionally relevant parts to provide a more intuitive emotion classification process. A feedforward neural network with one hidden layer with a final SoftMax classifier is used to classify 6 emotion-class probabilities from the emotion representation $e^{\sim}{}_t$ derived via attention mechanism for each respective utterance u_t. Here, $W_\beta$denotes a weight vector for the softmax layer and β_t is the attention score over the previous emotion states.

$$e_t = GRU_{emotion}(e_{t-1}, q_{s(u_t),t}) \tag{7}$$

$$\beta_t = softmax\left(e_t{}^T W_\beta \left[e_1, e_2 \ldots e_N\right]\right) \tag{8}$$

$$e^{\sim}{}_t = \beta_t [e_1, e_2 \ldots e_N]^T \tag{9}$$

3.3.5 Real-time emotion recognition

A conversation proceeds with t number of sessions shared among n number of participants, along with the utterance sequence $\{u_1,u_2,..,u_t\}$. The emotion corresponding to the utterance at index t is queried using historical utterances of the conversation up to the timestamp t. The prediction time is crucial for the real-time application and the performance measurement of the proposed approach is explained in the experiment section.

4 Experiments

The proposed emotion detection pathway is evaluated using two datasets: Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset and the Multimodal Emotion Lines Dataset (MELD) datasets of IEMOCAP and MELD. Table 2 shows the distribution of train and test samples for both datasets. For the IEMOCAP dataset, the proposed emotion detection methodology is also evaluated by varying the number of emotions facilitating appropriate comparison with existing research studies. When evaluating performance for five emotions, the predictions of happy and excited emotions are combined to represent the happy-emotion

Table 2 Statistical information of the dataset

Dataset	Partition	Utterance count	Dialog count
IEMOCAP [10]	train+eval	5810	120
	test	1623	31
MELD [32]	train+eval	11,098	1153
	test	2610	280

category. Furthermore, when evaluating the model performance for four emotions, the angry and frustrated emotion predictions are combined to reproduce the angry-emotion category. With this experiment, we demonstrate the high representational strength of BoAW based feature embeddings proposed in our approach by comparing accuracies (weighted and unweighted) in classifying 6, 5 and 4 categorical emotions derived from human conversations. By comparing the results with text modality based approaches, we further justify the reduction of the accuracy gap between text modality and audio modality. The performance of the novel pathway is evaluated using the MELD for 7 categorical emotion prediction. In addition, the ability of the novel feature pipeline to provide real-time predictions for online conversations is evaluated.

4.1 Datasets

IEMOCAP IEMOCAP [10] dataset is collected at Signal Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC) which consists of videos of two-way conversations conducted for five sessions by ten unique speakers where two speakers contribute per session for a scripted or an improvisation act to evoke the specified emotions. The corpus contains dyadic conversations and the dialogues segmented into utterances. Each utterance is labelled by 3 human annotators using categorical and dimensional (arousal and valence) labels and the majority voted emotion is taken as the label. We use 6 of the categorical labels in our study (happy, sad, neutral, angry, excited, frustrated) and train the model in a speaker-independent manner, i.e., the first four sessions (8 speakers) belong to the training partition while the last session (2 speakers) is taken as the testing partition. The residual utterances having miscellaneous labels are removed from the dialogue by assuming that the context of the conversation is not adversely impacted as 75% of the utterances fall under the six basic emotion categories.

MELD The MELD [32] is provided with multiple modalities and consists of multiparty conversations extracted from the Friends TV series for emotion detection tasks. Compared to dyadic conversations, multiparty conversations are more challenging for emotion detection. The majority voting of five annotators was taken by looking at the video clip of each utterance. The utterances are labelled with the 7 emotion classes according to Ekman's six universal emotions (fear, anger, surprise, sadness, joy and disgust) [15] extended with neutral emotion. The four emotions i.e., happy, sad, angry, neutral cover 96% percent of the emotion distribution in the training partition.

4.2 Experiment setup

We conduct experiments mainly on the IEMOCAP dataset and conduct an additional evaluation of the proposed approach using the MELD, despite a few facts related to its inherent weaknesses including constant background noise (laugh) and unrealistic, swift emotion switching. The audio files of the given utterances in the IEMOCAP dataset are directly used as input to the proposed pipeline. As only the conversational videos are present in the MELD, related audio streams were extracted from the videos and converted to the audio format (of 16-bit PCM WAV files) required by the openSMILE tool using ffmpeg [13]. The remainder of the experiment setup is common to both the datasets.

By specifically using the openSMILE configuration file *ComParE_2016*, for each speech of duration 25 milliseconds, 130 LLDs are extracted with a frame rate of 10ms. The low-level patterns are extracted from 10 milliseconds of speech segments in which the audio-related properties are assumed to be stationary. The utterances have variable lengths with an average time of 4.5 s. The applied bag of audio word approach is able to generate fixed size feature encodings by generating a term frequency matrix by taking the audio features into account irrespective of the duration of the utterance. The openXBOW open-source bag-of-words toolkit is a Java application which supports the generation of bag-of-audio-word representations from numerical feature sequences. We have used openXBOW toolkit to generate a 'bag of audio words' representation by providing the acoustic LLDs extracted from the audio data. The LLD feature vectors of the train partition are split into two sets to create two codebooks with 1000 indices yielding 2000 indices in total. Each codebook acts as the vocabulary for train and test feature vectors of 65 LLDs. A given 10-millisecond frame of an utterance, represented by a 130 length LLD vector is compared against existing patterns in the codebook by calculating Euclidean distances and the term frequency (TF) of the highest matching pattern's (least distant pattern) index is increased. In our approach, we use 5 as a parameter for the number of index matchings to be considered when comparing with the respective feature vector and update the indices of 5 highest similar term frequencies in the matrix. At the end of the process feature embeddings of 2000 dimensions are obtained for each utterance. The term frequency of the matrix of size 2000 is log-transformed before feeding to the emotion detection model.

In our experiment, the dimensions inside RNN are as follows: encoded utterance $u_t \in R^{2000}$, the global state of the conversation $g_t \in R^{150}$, speaker's state $p_t \in R^{150}$and emotional state of the speaker $e_t \in R^{100}$. In the emotion detection model, the Negative log-likelihood loss with L2 regularization is used and weights are assigned to each class to compensate for the data imbalance in the training partition. Stochastic gradient descent based Adam optimizer is used for optimization with a learning rate of 0.0001 and weight decay of 0.0001. The model is trained for 60 epochs with a batch size of 2 dialogue sequences.

4.3 Results

First, we present the results obtained in the IEMOCAP dataset with a comparison to the available state-of-the-art emotion detection models. Second, the result achieved from the experiment on MELD is discussed. Since the datasets are imbalanced, we measure the overall accuracy (weighted accuracy, WA) as well as average recall (unweighted accuracy, UA) over the different emotional categories.

IEMOCAP Our new approach resulted in achieving a weighted accuracy of 60.87% and an unweighted accuracy of 60.97%. To the best of our knowledge, this is the highest reported result for the 6 basic emotion classification using the audio modality for the IEMOCAP dataset. As shown in Table 3, we evaluated the performance improvements gained by the integrations of the proposed components, in particular the BoAW and attention mechanism. Initially, we construct a baseline emotion detection pipeline by extracting 6373 features from each utterance using *IS13 ComParE* configuration script available in openSMILE and feeding it to the BiDialogRNN emotion detection model (Baseline). Then the performance of the proposed BoAW integration to the BiDialogRNN is evaluated, which outperforms the baseline by 10% showing the applicability of the novel feature pipeline. This architecture is further enhanced by the proposed attention mechanism and this outperforms the baseline by 13%. The confusion matrix shown in Fig. 4

Table 3 Accuracy comparison of the proposed feature pipeline variants with the baseline

Approach	Weighted Accuracy	Unweighted Accuracy	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1 Score
Baseline	53.91%	49.14%	55.16%	53.91%	51.81%
BoAW+BiDialogRNN	59.64%	53.91%	57.72%	59.64%	57.69%
BoAW+BiDialogRNN+ Attention mechanism	**60.87%**	**60.97%**	**62.89%**	**60.87%**	**61.01%**

The significance of bold values is to emphasize the high accuracies obtained from our model comapred to the baseline model, in every accuracy category.

shows the performance of the proposed novel feature pipeline of emotion detection. The confusion matrix indicates that the misclassifications occur often among similar emotion classes which can be misinterpreted even in human-intervened classification. A set of happy emotions are misclassified as excited as the two emotions are close in nature. Some of the frustrated emotions are misclassified as angry as the roots for both the emotions are similar. Furthermore, most of the other emotion categories are frequently misclassified and tagged as neutral. The naive reasoning would be the centred-location of the neutral emotion on the activation-valence space [41] which attracts other emotions which are not clearly separated in terms of deviation on either side (positive, negative).

We evaluated the performance of the proposed novel approach, which utilizes the Bag-of-Audio-Words (BOAW) embeddings as input to a BiDialogRNN (with attention) model by comparing major state-of-the-art models in Table 4. As most of the emotion detection models are limited to a few basic emotions including happy, angry, neutral and sad, we aggregate happy with excited and angry with frustrated for convenient comparison with models which cater to a lower number of emotions. The work in [20] proposed a multimodal emotion detection framework updating only the global conversational context could achieve good accuracy, but the performance of the audio modality is lower compared to its text modality. The model fusion of audio and text for identifying 4 emotions in [40] has improved the weighted accuracy to 76%, but the unweighted accuracy of 69% is still less than our results. The work in [29] proposed RNN-weighted pool approach with

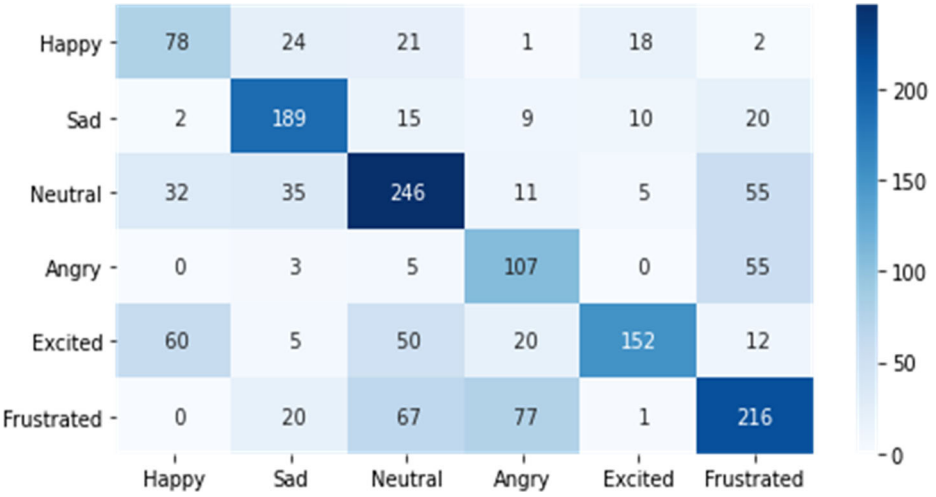

Fig. 4 Confusion Matrix for Emotion detection on IEMOCAP Dataset

Table 4 Accuracy Comparison for IEMOCAP dataset

Proposed Approach	Model	Number of Emotions	Weighted accuracy (%)	Unweighted accuracy (%)
[20]	ICON	6	50.7%	-
Ours	BoAW + BiDialogRNN + Attention		**60.87%**	**60.97%**
[18]	DNN + ELM	5	54.3%	48.2%
Ours	BoAW + BiDialogRNN + Attention		**65.6%**	**66.07%**
[40]	CNN + DNN	4	73.6%	62.9%
[29]	RNN + attention		63.5%	58.8%
[41]	Audio Recurrent Encoder		-	54.6%
[42]	MULTI-HOP ATTENTION Model		64.6%	65.2%
[11]	3-D Convolutional Recurrent Neural Networks with Attention Model		-	64.74%
Ours	BoAW + BiDialogRNN + Attention		**73.81%**	**73.32%**

The significance of bold values is to emphasize the high accuracies obtained from our model comapred to the baseline model, in every accuracy category.

attention, which improves their accuracy. As most of the reported work is based on utterance level predictions without considering contextual information of the conversation, it highlights the importance of conversational modeling. It can be seen from the comparison that the proposed combination of bag-of-audio-words audio embedding and the recurrent network yields state-of-the-art results.

MELD Compared with the IEMOCAP dataset, MELD dataset is more challenging. One reason is that the average number of speaker sessions is 10 (shorter conversations) in the MELD dataset, whereas it is 50 in the IEMOCAP dataset. In addition, there exists a rapid emotion switching with an average of three emotion categories per dialogue in MELD which adversely affects the attention-based context capturing mechanism in the RNN. This issue is exacerbated by having more than 5 speakers present in the majority of the conversations resulting in less information to keep track of the speaker states. In addition, the average length of an utterance is lower (3.59 s) compared to the IEMOCAP dataset (4.59 s), resulting in reduced emotion detection ability. Although all seven emotion categories are input for the training phase, the dataset is severely imbalanced with 47% of neutral data while the fear and disgust classes include 2% of the training samples. The background noise of laughing in a majority of utterances leads to increased incorrect classifications. This has resulted in MELD dataset is hardly used for emotion classification with audio in recent literature. Thereby, in our work, we compared our audio based emotion classification approach with recent text based emotion classification approaches with MELD dataset. Comparison of accuracy with the baselines provided by MELD dataset are shown in Table 5. As such, we achieved the highest F1-score of identifying disgust, joy and sadness emotions outperforming the baselines provided by the

Table 5 Accuracy Comparison for MELD dataset

Model	Source	F1 - Score						
		Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
bcLSTM [32]	Text	25.85	6.06	2.90	15.74	61.86	14.71	19.34
DialogueRNN [32]	Text	**35.18**	5.13	**5.56**	13.17	**65.57**	14.01	**20.47**
Ours	Audio	28.65	**6.50**	2.17	**17.16**	38.79	**15.24**	13.38

The significance of bold values is to emphasize the highest accuracies obtained from various models, in each emotion category.

MELD dataset. Due to the limitations highlighted, emotion detection accuracies are lower than other IEMOCAP dataset. Considering the comparable performance with the state of the art, we believe it is important to present these outcomes as these will inform interested researchers to further explore and take the work forward.

4.4 Comparison with text based emotion detection

The results from the MELD dataset are compared with the text modality based existing state-of-the-art emotion models in order to highlight the representational ability and the robustness of the audio embeddings of the proposed pipeline. The previous attempts for emotion detection using the textual features have reported higher accuracies compared to audio feature feature-based techniques. The proposed model successfully utilizes audio features to predict 6 basic emotions for the MELD dataset achieving comparable average accuracy and F1 scores (Table 6) achieving on par accuracies with textual modality based state-of-the-art models.

4.5 Impact of proposed attention mechanism

The attention mechanism is used to capture the emotionally relevant segments of the conversation to enrich the emotion detection process. When predicting the emotion for each utterance, in addition to the utterance, weighted conversation-contexts from the preceding utterances are considered by means of attention scores. The current global state of the conversation is the query parameter to retrieve the attention scores. In order to elaborate on the impact of the proposed attention mechanism, the utterance sequence of the dialog "Ses05F_impro03" in the IEMOCAP test partition is examined. This is a conversation between two parties, continuing an excited and happy emotion mixed conversation in which one person initially announces her marriage proposal. The excitement generated from this major occurrence at the beginning of the conversation continues forward during the conversation. Figure 5 illustrates snapshots of attention scores given by both the forward and backward RNNs of the model, which extract information from the past and future states of the conversation respectively. The attention scores elicited at the presence of the utterance at index 20 in the conversation is shown in Fig. 5(a) which indicates that the utterances residing in the range 5 to 12 have gained more attention indicating the initial excited state of the actual conversation. When predicting emotion for the utterance at the index 40 of the conversation (Fig. 5(b)), it denotes that attention has been influenced by not only the neighbouring utterances in the index range from 24 to 30, but also the significant incident captured in the 4 to 12

Table 6 Performance comparison with text based models on IEMOCAP

Model	Source	Happy	Sad	Neutral	Angry	Excited	Frustrated	Average	
		Acc	Acc	Acc	Acc	Acc	Acc	Acc	F1
c-LSTM [31]	Text	30.5	56.7	57.5	59.4	52.8	65.8	56.3	56.1
CMN [19]	Text	25	55.9	52.8	61.7	55.5	**71.1**	56.5	56.1
Dialogue RNN [28]	Text	25.6	75.1	58.5	64.7	**80.2**	61.1	63.4	62.7
Dialogue GCN [17]	Text	40.6	**89.1**	61.9	**67.5**	65.4	64.1	**65.2**	**64.1**
Ours	Audio	**54.2**	77.1	**64.06**	62.94	50.84	56.69	60.9	61

The significance of bold values is to emphasize the highest accuracies obtained from various models, in each emotion category for the text modality.

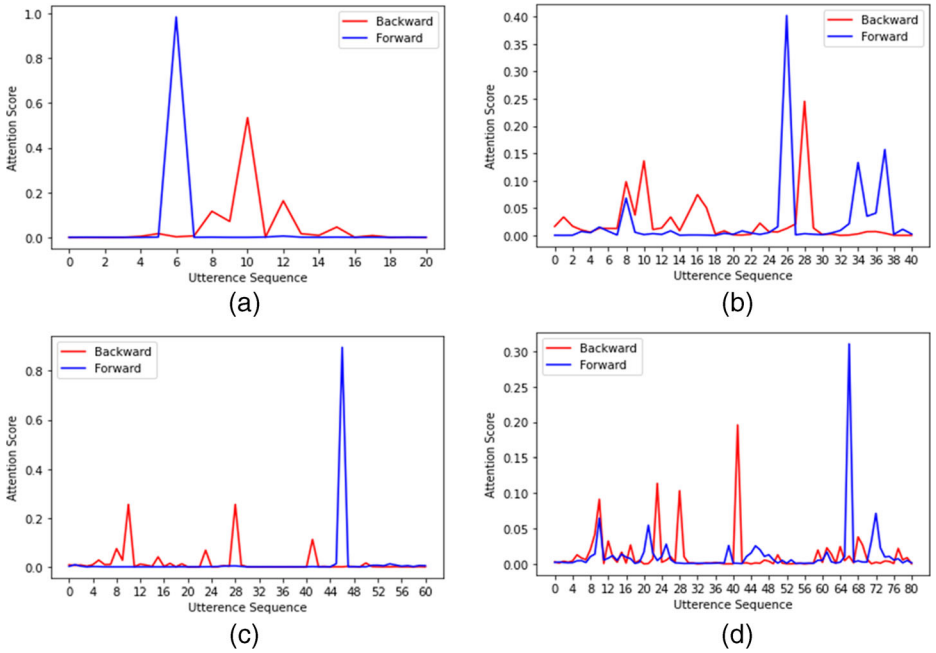

Fig. 5 The attention scores for the preceding utterances at each quarter of the conversation "Ses05F_impro03", taken from the IEMOCAP test set. (**a**) After 20 utterances, (**b**) after 40 utterances, (**c**) after 60 utterances and (**d**) after 80 utterances of the conversation

utterance range. Similarly, the residual plots for the utterances at indices 60 and 80 of the conversation (Fig. 5(d)) preserve the context. The illustration demonstrates that the attention mechanism has the ability to capture the emotionally salient segments, unlike in the independent utterance-wise emotion detection methodologies. The proposed attention mechanism provides better interpretability and context preservation to the emotion detection model.

4.6 Real time emotion detection

Latency during real-time emotion prediction is measured by deploying the proposed emotion detection approach. The sub-modules of the emotion detection approach are tested separately, and the results shown in Table 7 show the applicability of the proposed approach in real-time emotion prediction settings. The evaluation is conducted for the test partition of the IEMOCAP dataset, including 31 dialogs consisting of 1623 utterances. At each session of the speaker utterance, the openSMILE toolkit extracts features and generates audio-relevant embeddings via the openXBOW tool. For the task of emotion detection, with the audio encoding of the current utterance, the previous speaker utterances are also provided as input to the model in order to yield context-influenced emotion predictions. As a byproduct, the time consumed for deriving emotion predictions increases with the length of the conversation, i.e., the time for retrieving the emotion prediction for the final utterance is equal in value to the time to retrieve the emotion sequence of the full conversation. As IEMOCAP dataset contains 50 utterances in average per conversation, in the worst possible scenario it is justified to assume the time taken for retrieving an emotion for an arbitrary utterance, is the same as the time spent to retrieve

Table 7 Real-time performance statistics

Phase	Time spent (milliseconds)			
	Total time (1623 utterances)	Average time per utterance	Average time per conversation	Real-Time task latency (seconds)
Feature Extraction	553,443	341.21	-	0.341
Audio embeddings	213,826	131.747	-	0.137
Predict Emotion	6,460	-	208	0.208
Total				**0.686**

The significance of the [bold] is to emphasize the overall time duration for the feature extraction, audio embedding and emotion prediction steps of the pipeline.

emotion predictions for a conversation having 50 utterances on average. The results suggest approximately 0.7 s of latency per utterance for passing through the complete pipeline.

5 Discussion and conclusion

Emotion detection from human conversations in audio form is a key challenge which can provide significant benefits if successfully overcome. The proposed research has contributed multiple theoretical contributions to the research community. Although the Bag-of-Audio-Words (BoAW) based embeddings have been used for the detection of the arousal and valence of emotions, to the best of our knowledge, this is the first study which highlights the potential of the BoAW based feature representations for basic and complex human emotion detection tasks. The evaluation of the proposed novel feature pipeline conducted upon the IEMOCAP dataset yields promising results of 60.87% weighted accuracy (approximately 20% improvement) and 60.97% unweighted accuracy in recognizing the 6 basic emotions outperforming current state-of-the-art models using the audio modality. In a multi-modal setting where audio, text and video streams are present, emotion recognition can be done with higher accuracy due to the richness of a wide variety of features. However, when only the audio modality is available for analysis, such as in a customer care call centre, presence of an emotion detection modal which is trained only using audio modality and can provide reasonable level of accuracy is very useful. Although the accuracies shown are approx. 61%, these far exceed the current state of the art (by 20%) highlighting the potential of the novel emotion recognition pipeline for audio/acoustic data. As potential future work, the proposed pathway can be experimented on tracking emotions of multiparty scenarios as the adapted emotion detection model is scalable enough to be evaluated for multiple parties. Going beyond the basic emotion categories, the research can be extended to detect mixed emotions by adding a final mixed-emotion classifier by utilizing the probability values yielded from the proposed RNN model [5]. This mixed-emotion classification can benefit from the theoretical backgrounds of the plutchik's emotion wheel [30]. As another future direction, the feature extraction can be strengthened by augmenting the extracted auditory features by textual features from Bag-of-Words (BoW) by utilizing automatic speech detection techniques resulting in the potential utilization of big audio data. Beyond the traditional BoAW based feature encoding, the Convolution Neural Network-based deep learnt feature extraction techniques utilizing sequential information such as wave2vec 2.0 have emerged in the arena of speech recognition which could be further experimented on downstream tasks including emotion recognition from human conversations [9]. During the research, a limiting factor identified for deploying the modal in a real-world setting is the lack of devised state-of-the-art

solutions for real-time automatic speaker diarization. With the availability of such speaker diarization, a fully automated emotion recognition pipeline from human conversations could be developed based upon the research contributions described in this paper. It is anticipated such input the accuracy will improve significantly. The proposed feature pipeline is simulated for real-time emotion detection from human conversations via manually diarized conversations due to the non-existence of a robust real-time technique. The proposed system can be implemented with high commercial use by integrating with a real-time speaker diarising methodology. A mechanism to generate speech embeddings from the bag of audio words from a large corpus of speech data can be more reliable as the bag of audio word encodings have a high dimension with high sparsity where the generated embeddings represent rich information.

The proposed research contributes a number of practical innovations to the modern tech-driven world. If machines become capable of identifying complex human emotions, various systems including elderly care agents, virtual call centre agents and specially designed AI robots can provide an improved customized service. Organizations which consider customer satisfaction as a major driving force of their businesses can benefit by analyzing how the agents manage the customers via exploring the emotion variations throughout the agent-customer conversations. Although understanding emotions from daily conversations is natural for human beings, the detection of emotions only through audio conversations could be difficult without use of facial expressions. In this context, empowering machines to understand human emotions via audio conversations is a significant step advancing the field of human-computer interaction by enhanced leveraging of big audio data.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflicts of interests The Authors declare that there is no conflict of interest.

References

1. Abeysinghe S et al. (2018) Enhancing decision making capacity in tourism domain using social media analytics. 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), pp 369–375. https://doi.org/10.1109/ICTER.2018.8615462
2. Adikari A, Alahakoon D (2021) Understanding citizens' emotional pulse in a smart city using artificial intelligence. IEEE Trans Ind Inf 17(4):2743–2751. https://doi.org/10.1109/TII.2020.3009277
3. Adikari A, Burnett D, Sedera D, de Silva D, Alahakoon D (2021) Value co-creation for open innovation: An evidence-based study of the data driven paradigm of social media using machine learning. Int J Inf Manag Data Insights 1(2):100022

4. Adikari A, Nawaratne R, De Silva D, Ranasinghe S, Alahakoon O, Alahakoon D (2021) Emotions of COVID-19: Content analysis of self-reported information using artificial intelligence. J Med Internet Res 23(4):e27341
5. Adikari A, Gamage G, de Silva D, Mills N, Wong S, Alahakoon D (2021) A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web. Futur Gener Comput Syst 116:302–315
6. Alahakoon D, Nawaratne R, Xu Y, De Silva D, Sivarajah U, Gupta B (2020)Self-building artificial intelligence and machine learning to empower big data analytics in smart cities. Inform Syst Front. https://doi.org/10.1007/s10796-020-10056-x
7. Alvi S, Afzal B, Shah G, Atzori L, Mahmood W (2015) Internet of multimedia things: Vision and challenges. Ad Hoc Networks 33:87–111
8. Arthur D, Vassilvitskii S (2007) k-means++: The advantages of careful seeding. In: Proc. of the 18th annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, pp 1027–1035
9. Baevski A, Zhou H, Mohamed A, Auli M (2021) wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv.org
10. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: Interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335
11. Chen M, He X, Yang J, Zhang H (2018)3-D Convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process Lett 25(10):1440–1444. https://doi.org/10.1109/LSP.2018.2860246
12. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. https://doi.org/10.48550/arXiv.1412.3555
13. Converting Video (2020) Formats with FFmpeg | Linux Journal. Linuxjournal.com
14. Devamanyu Hazarika S, Poria A, Zadeh E, Cambria L-P, Morency, Zimmermann R (2018) Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume1 (Long Papers), vol 1, pp 2122–2132
15. Ekman P (1992) An argument for basic emotions. Cognit Emot 6(3–4):169–200. https://doi.org/10.1080/02699939208411068
16. Florian Eyben F, Weninger F, Gross B (2013) Schuller: Recent Developments in open SMILE, the Munich Open-Source Multimedia Feature Extractor. In: Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp 835–838. https://doi.org/10.1145/2502081.2502224
17. Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A (2019) Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. arXiv preprint arXiv:1908.11540
18. Han K, Yu D, Tashev I (2020) Speech emotion recognition using deep neural network and extreme learning machine. Microsoft Research
19. Hazarika D, Poria S, Zadeh A, Cambria E, Morency L-P, Zimmermann R (2018) Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1 (Long Papers), pp 2122–2132
20. Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R (2020) ICoN: Interactive conversational memory network for multimodal emotion detection. Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018, pp 2594–2604. https://doi.org/10.18653/v1/d18-1280
21. De Barros PVA (2016) Modeling affection mechanisms using deep and self-organizing neural networks. Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky
22. Izard C (2013) Human emotions. Springer, New York, pp 1–4
23. Jiao W, Lyu MR, King I (2019)Real-time emotion recognition via attention gated hierarchical memory network. arXiv preprint arXiv:1911.09075
24. Keren G, Schuller B (2016) Convolutional RNN: An enhanced model for extracting features from sequential data. Proc. Int. Jt. Conf. Neural Networks, vol. 2016-October, pp 3412–3419. https://doi.org/10.1109/IJCNN.2016.7727636
25. Lee C-C, Mower E, Busso C, Lee S, Narayanan S (2011) Emotion recognition using a hierarchical binary decision tree approach. Speech Commun 53(9–10):1162–1171
26. Lieskovská E, Jakubec M, Jarina R, Chmulík M (2021) A review on speech emotion recognition using deep learning and attention mechanism. Electronics 10(10):1163
27. Madhavi I, Chamishka S, Nawaratne R, Nanayakkara V, Alahakoon D, De Silva D (2020) A deep learning approach for work related stress detection from audio streams in cyber physical environments. 2020 25th IEEE International Conference on Emerging Technologies and Automation F (ETFA), pp 929–936. https://doi.org/10.1109/ETFA46521.2020.9212098

28. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E (2019) DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 6818–6825. Available: https://doi.org/10.1609/aaai.v33i01.33016818
29. Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention center for robust speech systems. The University of Texas at Dallas, Richardson, TX 75080, USA Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA. IEEE Int. Conf. Acoust. Speech, Signal Process, pp 2227–2231. https://doi.org/10.1109/ICASSP.2017.7952552
30. Plutchik R (2001) The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am Sci 89(4):344–350
31. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P(2017)Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (vol 1: Long Papers), pp 873– 883
32. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R (2019) MELD: A multimodal multi-party dataset for emotion recognition in conversations. ACL, pp 527–536
33. Rathnayaka P, Abeysinghe S, Samarajeewa C, Manchanayake I, Walpola M, Nawaratne R, Bandaragoda T, Alahakoon D (2019) Gated recurrent neural network approach for multilabel emotion detection in m-icroblogs. 2012:2012–2017. http://arxiv.org/abs/1907.07653
34. Rosalind WP (2010) Affective computing: from laughter to IEEE. IEEE Trans Affect Comput 1(1):11–17
35. Ruusuvuori J (2013) Emotion, affect and conversation. The handbook of conversation analysis, pp 330–349
36. Satt A, Rozenberg S, Hoory R (2017) Efficient emotion recognition from speech using deep learning on spectrograms,. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol 2017-August, pp 1089–1093. https://doi.org/10.21437/Interspeech.2017-200
37. Schmitt M, Schuller B (2017) openXBOW - Introducing the passau open-source crossmodal bag-of-words toolkit. J Mach Learn Res 18(96):1–5
38. Schmitt F, Ringeval, Schuller B (2016) At the border of acous-tics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. Proc of Interspeech, pp 495–499
39. Schuller B, Steidl S, Batliner A, Epps J, Eyben F, Ringeval F, Marchi E, Zhang Y (2014) The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In: Proceedings INTERSPEECH 2014. 15th Annual Conference of the International Speech Communication Association, (Singapore, Singapore), ISCA, ISCA
40. Tripathi S, Kumar A, Ramesh A, Singh C, Yenigalla P (2019) Deep learning based emotion recognition system using speech features and transcriptions, pp 1–12
41. Yoon S, Byun S, Jung K (2019) Multimodal speech emotion recognition using audio and text. 2018 IEEE Spok. Lang. Technol. Work. SLT 2018 - Proc., no. December, pp 112–118. https://doi.org/10.1109/SLT.2018.8639583
42. Yoon S, Byun S, Dey S, Jung K (2019) Speech emotion recognition using multi-hop attention mechanism. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2822–2826

Affiliations

Sadil Chamishka [1] · Ishara Madhavi [1] · Rashmika Nawaratne [2] · Damminda Alahakoon [1] · Daswin De Silva [2] · Naveen Chilamkurti [3] · Vishaka Nanayakkara [1]

Sadil Chamishka
sadilchamishka.16@cse.mrt.ac.lk

Ishara Madhavi
madhavi.16@cse.mrt.ac.lk

Rashmika Nawaratne
B.Nawaratne@latrobe.edu.au

Damminda Alahakoon
D.Alahakoon@latrobe.edu.au

Daswin De Silva
D.DeSilva@latrobe.edu.au

Vishaka Nanayakkara
vishaka@cse.mrt.ac.lk

[1] Computer Science and Engineering, University of Moratuwa, Moratuwa, Sri Lanka

[2] Research Centre for Data Analytics and Cognition, La Trobe University, Victoria, Australia

[3] Computer Science and Computer Engineering, La Trobe University, Victoria, Australia