MULTIMODAL EMOTION RECOGNITION USING DEEP LEARNING TECHNIQUES

Enguerrand BOITEL, Alaa MOHASSEB, Ella HAIG

Abstract—Human emotion recognition is a topic that interests many deep thinkers in the field of artificial intelligence nowadays. Indeed, there have been significant advancements in this sector in the last few years. Emotions are expressed through voice, hand or even body gestures and facial expressions. For instance, the expressions on a person's face reveal a lot about their cognitive process and provide insights into what is happening inside their heads. The goal of emotion recognition is then to familiarise the computer with the ability to perceive human emotions in the same way that humans do. A lot of work has been done regarding emotion recognition for specific data, for example, text (using NLP, Text Emotion Recognition - TER), audio (Speech emotion recognition - SER) or image (Face Emotion Recognition - FER), but there are just a few papers related to Motion Emotion Recognition - MER. Therefore, the goal of this research is to categorise four different data streams (text, audio, images and videos) into one of the seven main emotions, recognised as basic emotions using deep learning methods.

Index Terms—Deep Learning, Emotion Recognition, Natural Language Processing, Neural Networks

1 INTRODUCTION

EMOTIONS are spontaneous, unconscious psycho-physiological mechanisms that can be influenced by a variety of elements (e.g. attitude, personality). Indeed, emotions play a significant role in human judgement, which makes their study crucial. In fact, Automatic Emotion Recognition systems are becoming more and more necessary as technology develops and our comprehension of sentiments increases. Consequently, text (TER), speech (SER) and facial expressions (FER) has been used extensively in emotion recognition. The BAUM (Bahçeşehir University Multimodal Affective Database) dataset is one of the largest publicly available multimodal resources for emotion recognition. It is made up of about 1200 facial video clips from 31 subjects, including speech records and text translations. Also, the SAVEE (Surrey Audio-Visual Expressed Emotion) database, containing 480 video clips from 4 actors, was utilised in this research. Sentences from the common TIMIT corpus were chosen to create the dataset, and phonetically balanced for each mood.

Therefore, in this study, a combination of these 4 modalities is proposed, to create a more powerful and reliable emotion recognition system. In order to obtain the greatest individual detection accuracy from each stream, several deep learning-based architectures were first investigated. Then, each stream using the results of the more effective models were combined. This investigation has two benefits. First, because each stream was focused independently, the lack of accuracy of one modality does not cripple the algorithm. This enables the proposed strategy to be modular as well. Second, videos were used to recognise emotions through motion, and subtitles to detect sentiments through text, two reliable outputs. Additionally, cutting-edge hyper-parameter optimisation algorithms to find the optimum model configuration feasible under the restrictions of available resources were developed.

2 RELATED WORKS

Concerning text, authors in [1] suggested employing a deep learning model at the SemEval 2018 Task 1 Competition ([2]). They designed a multi-layer self-attention system incorporated into a Bidirectional Long-term Short-Term Memory (Bi-LSTM). According to the experimental findings, random initialisation outperformed this transfer learning approach. Moreover, a supervised logistic regression method was revealed in [3] to identify emotions in text. Sentences with emotions were supplied into a logistic regression model and emotion labels throughout the training procedure. In their testing process, the trained classifier was only fed texts with hidden emotion labels. Precision, recall, and F1-score were used to assess the effectiveness of their model. For the feelings studied (joy, fear, sadness, guilt and shame), they obtained an F1 score between 0.57 and 0.76. Over 144,000 tweets were used in [4] to construct an emotion recognition model using CNN. The emotion of *happiness* was identified with the highest accuracy rate (73%), whereas *anger* had the lowest precision level (37%). Also, to automatically forecast the authors' personality sentiments, authors in [5] derived contextualised word embeddings from text data using the BERT pre-trained model and the bagged-SVM classifier. Comparing their performance to baseline methods, they obtained an increase of 1%.

The early studies on SER employed hand-crafted speech characteristics for classification. Based on basic amplitude and tone, [6] extracted speech features and assessed how they related to different emotions. Over the years, a number of algorithms to identify feelings in human speech have been suggested. Algorithms for machine learning including Support Vector Machines (SVM - [7]), Hidden Markov Models (HMM - [8]), Gaussian mixture models (GMM - [9]), etc. were frequently presented. Also, CNNs have been

utilised by [10] to identify voice emotions. Finally, [11] have proposed a very effective bidirectional Recurrent Neural Network (RNN) at extracting crucial speech features for improved SER performance.

Regarding FER, [12] demonstrated the CNN' capability in the emotion recognition tasks in comparison to the conventional SVM (Support Vector Machine), for the two databases CK+ [13] and FER 2013 [14]. However, overfitting can occur when CNN is directly trained on a tiny facial expression database. Also, [15] demonstrate how improving performance and addressing the issue of insufficient training data may be accomplished by fine-tuning a pre-trained model using the second dataset of facial movements. Additionally, [16] reported that performing FER along with other tasks, including landmark localisation, might enhance its performance.

In earlier MER research, the optical flow feature has most frequently been utilised. For example, from RGB images and optical flow images, [17] used two CNNs to extract features. Two different LSTMs were fed with the collected features to forecast emotions.

Compared to unimodal learning, the multimodal study is significantly more effective (as presented in [18]). Studies have also attempted to combine signals from several modalities, such as facial expressions and audio, speech and text, and various combinations of these modalities, for increased effectiveness and precision. Currently, this method has been expanded to further improve the accuracy of emotion recognition. The multimodal fusion model can detect feelings by fusing physiological inputs ([19]) in a variety of ways. Deep learning architectures have recently undergone advancements, and this has led to its application in multimodal emotion recognition. On the subject of classifying audio-visual communication, [20] reported remarkable findings. They employ Restricted Boltzmann Machines (RBMs) for multimodal fusion. Kahou and his associates performed emotion identification from video clips using a collection of deep learning models [21]. They won the Emotion Recognition in the Wild Challenge [22] with this. Likewise, Lee *et al.* ([23]) created a multimodal deep learning model that uses verbal descriptions of the situation and photos of faces to represent facial expressions. They designed two models to recognise emotions using images and texts. The results of the experiment showed that the effectiveness of emotion recognition is greatly increased when text definitions of their behaviour are used.

3 Experimental setup

3.1 Dataset

The BAUM dataset is composed of 1184 videos from 31 subjects, which resulted in more than 3 hours of audio-visual recordings. The video clips were created by taking frontal and half-profile views of the subjects with stereo and mono cameras, respectively. A series of images and a short video clip were first shown to the subjects, which have been selected to inspire different emotions and mental states. Secondly, the actors present their thoughts and feelings regarding the pictures and videos they have just viewed. The six fundamental emotions according to Ekman [24] (anger, disgust, fear, happiness, sadness and surprise) labelled the recordings, as well as six emotional states (boredom, bothered, concentrating, contempt, thinking and unsure). The "neutral" expression was finally added. Each expression is distributed, after normalisation (based on Ekman's emotions) as presented in Table 1.

Emotion	Number of recordings	Distribution
Anger	95	8%
Happiness	273	23%
Sadness	237	20%
Fear	47	4%
Disgust	154	13%
Surprise	59	5%
Neutral	319	27%
Total	1184	100%

TABLE 1: Distribution of emotion classes in the BAUM dataset

3.2 Framework

A multimodal emotion recognition system is built using four modes (text - transcripts of the actor's speech, speech - sound extracted from the videos, images - frames and motion - facial movements), following diagram 1. From videos, the library "AudioSegment" from "pydub" [25] was used to extract the speech (".wav" format). Then, the Google "SpeechRecognition" library [26] was utilised to transcript the resulting speech into text. Finally, the recordings were split into multiple frames using "cv2" and its function "VideoCapture" [27]. Afterwards, each data stream was used by a specific emotion recognition model to identify the emotion felt by the actor. Ultimately, all the predicted probabilities for each emotion, within each data type, will be combined in a final system to obtain the resulting emotion.

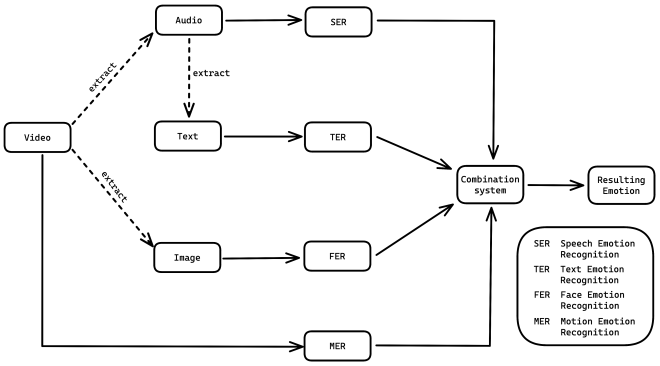
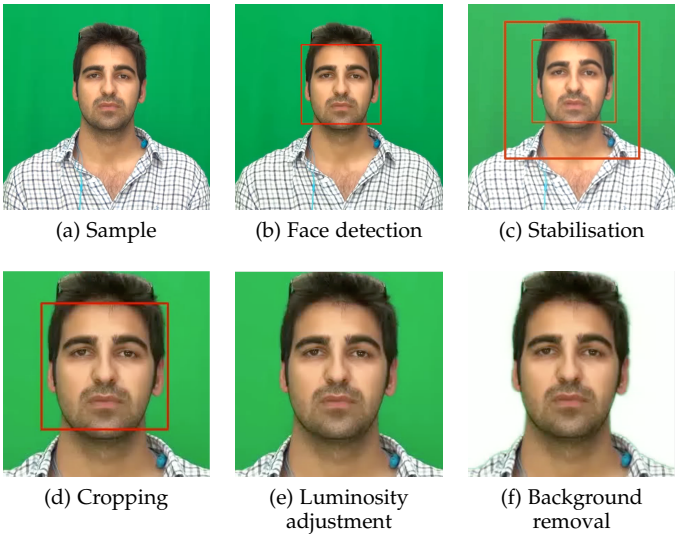


Fig. 1: Project framework

3.3 Data Preprocessing

Data in both BAUM and SAVEE datasets are in raw form and has to be processed to remove unnecessary image parts, audio or text. It involves some data mining techniques that enables the transformation of raw data into a usable and effective format. Preprocessing data involves the accuracy and efficiency of Deep Learning models.

Regarding images preprocessing tasks and by extension video (presented in Figure 2 using a BAUM sample - Figure



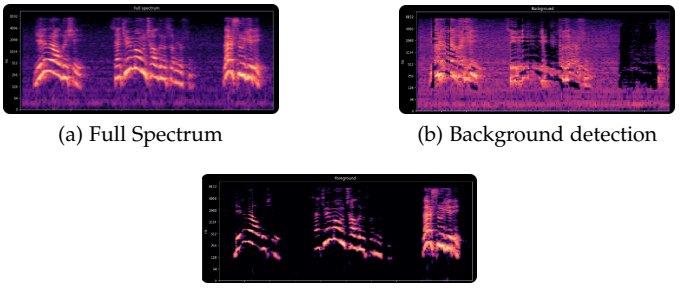
(a) Sample (b) Face detection (c) Stabilisation

(d) Cropping (e) Luminosity adjustment (f) Background removal

Fig. 2: Image preprocessing tasks: BAUM dataset - S001_001

2a), the position of the actor's face was first detected (Figure 2b) and then the video composed of the cropped images (Figure 2c was stabilised (Figure 3a). Finally, the luminosity was adjusted (Figure 3b) and the background removed (Figure 9a) in order to keep the best possible quality and avoid distraction for the deep learning model. Video preprocessing consists solely of the grouping of previously processed images.

For audio, the background was removed using the library "noisereduce" [28] (refer to Figure 3, before - 3a - and after - 9a - removing the noise - 3b) and the audio cut, so that only the sound levels where the main actor/subject is speaking, are taken into account. This was particularly interesting for FER, where the resulting emotion is averaged over all the frames in the video. It would indeed not be useful to keep images where the actor does not express any emotion. MFCCs (Mel-Frequency Cepstral Coefficients) features (inspired by [29]) and MFBs (Mel Filter Bank - [30]) were used to understand sound signals by extracting audio features.



(a) Full Spectrum (b) Background detection

(c) Clean audio

Fig. 3: Audio preprocessing tasks: BAUM dataset - S002_005

Concerning text, removing punctuation and stop words, lower casing, tokenising and stemming were some of the main preprocessing steps performed, which led the results from, for example, "*I am very curious... Come on please tell it. I promise...*" to "*[curi, com, promis]*" (using the Python's NLTK library, inspired by [31]).

4 Methods and Models Architecture

4.1 Text Based Emotion Recognition

BERT [32] is a language model that excels in various NLP tasks and is trained on a vast quantity of text data. The BERT model is dependent on the context of the sentence and learns contextual information from input texts. The BERT technique defeats word-based representations (Word2Vec - [33]) in many contexts, by utilising the sub-word representation. More specifically, the novel DeBERTa (Decoding-enhanced BERT with disentangled Attention) [34] significantly improves the efficiency of BERT by representing each word with two vectors that separately code its position and content. Also, the output softmax layer is replaced with an enhanced mask decoder in order to anticipate the masked tokens for the model pre-trained. This model was the first to surpass human performance on SuperGLUE.

In the proposed research, the DeBERTa-V3-Large pre-trained model was used, as presented in Figure 4. Unlabelled data from English Wikipedia (12GB), BooksCorpus (6GB), and Reddit content (38GB) were used to train this model. This model is composed of 24-layered transformer blocks where each block contains 1024 hidden outputs and has 304 million parameters in total. Additionally, randomly initialised fully connected layers were developed with 768×4 parameters along with softmax and fine-tune for emotion recognition on the BAUM using a categorical cross-entropy loss as the objective function, an Adam optimiser with a learning rate of $1{\times}10^{-5}$, and a batch size of 16. Finally, the PyTorch framework for the BERT model training was used.

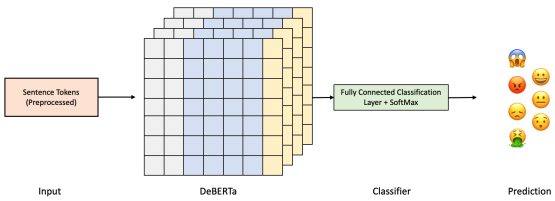


Fig. 4: DeBERTa framework for text-based emotion recognition

4.2 Audio Based Emotion Recognition

While many state-of-the-art SER systems have recently used complex neural network models that learn directly from spectrograms or even raw wave forms to capture and represent the various emotions in speech, SER systems have traditionally used a large set of low level time and frequency-domain features to do so. In this research, utilising MFCC spectrograms as input characteristics, an end-to-end CNN-based system was built. In recent years, speech applications like speaker recognition have taken an interest in CNN

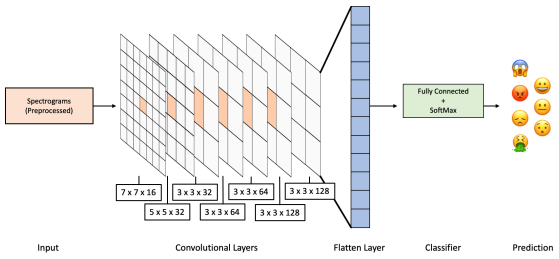


Fig. 5: CNN framework for audio-based emotion recognition

models that were initially created for SER and used for computer vision applications. To help the model learn the spectrum envelope and the coarse harmonic structures for the various emotions, high-resolution spectrograms were extracted.

Nine layers make up the proposed deep CNN, of which 7 convolutional layers and 2 fully connected layers are supplied to Softmax to generate probabilities of speech emotions. Convolutional filters are applied to the resulting spectograms as input to extract feature maps from a specific speech spectogram. Figure 6 presents the shape of each kernels. A 25% dropout ratio is included after the first completely linked layer to address overfitting. The use of strides, which eliminates the pooling layer while downsampling, and uses the same filter size and shape across the network to learn deep features, is the important element of this architecture. By utilising fewer fully connected layers, redundancy was minimised.

4.3 Image Based Emotion Recognition

A traditional neural network known as ResNet, or Residual Networks, serves as the foundation for many computer vision tasks. ResNet represented a significant advancement in that it effectively enabled to train deep neural networks with more than 150 layers. Also, it was the first to develop the concept of skip connections.

3x3 filters were utilised as the image is passed through a stack of convolutional layers. The spatial padding of the convolutional layer input is such that the spatial resolution is retained after convolution (i.e., the padding is 1 pixel for 3x3 convolutional layers), and the convolution stride is fixed to 1 pixel. Five max-pooling layers, which come after some of the convolutional layers, perform spatial pooling (max-pooling does not follow every layer). A 2x2 pixel window is employed to perform max-pooling.

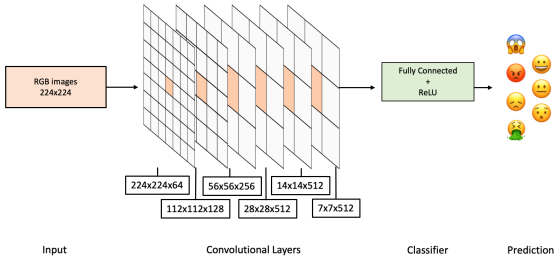


Fig. 6: ResNet-50 framework for image-based emotion recognition

4.4 Motion Based Emotion Recognition

Multiple hundreds (or thousands) of frames must be processed in order to assign a video to a specific mood. Nevertheless, the same video has very often identical facial expressions that are indicative of a specific emotion. Indeed, in order to effectively characterise and learn the emotion contained in an emotion video, a set of a few key-frames will be enough in terms of the variety of facial expressions (refer to Figure 7).

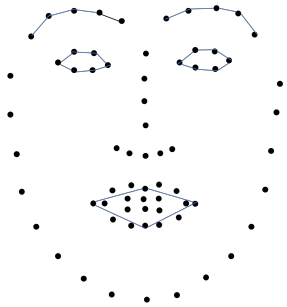


Fig. 7: Landmarks Recognition: SAVEE dataset - DC_h01

3DCNN is an extension of the conventional CNN with altered convolution and pooling methods. It is used to represent the spatio-temporal characteristics of lengthy sequences. The resilience of recognition algorithms may be impacted if dependencies between segments are ignored in long-duration sequences like speech, movies, and EEG signals. By performing 3D convolution operations on the input segments, the 3DCNN represents these temporal dependencies. Additionally, the 3D convolution technique can be used to visualise and model landmarks between video frame pixels.

Images of face expressions derived from video that are successively overlaid over five frames are the input data (Figure 8). Data size is initially supplied as 64x48x5. The size of the data changes to 60x44x3 at the second layer due to a convolution with a kernel that is 5x5x3. The third layer's data size changes to 30x22x3 due to a subsampling operation using a kernel with a size of 2x2x1 without a time base. The fourth layer's data size is the modified to 26x18x1 due to a convolution operation using a 5x5x3 kernel. Finally, the fifth layer's data size moved to 13x9x1 due to a subsampling operation using a kernel with a size of 2x2x1 without a time base.

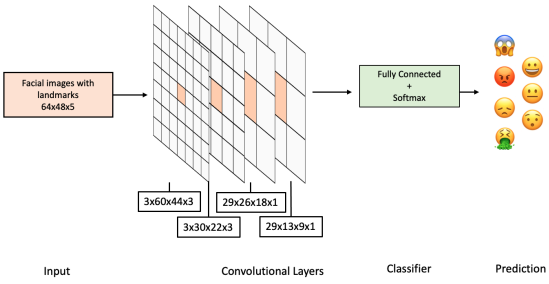


Fig. 8: 3DCNN framework for motion-based emotion recognition

Regarding the classification process, the Euclidean distance, following Equation 1, was utilised. The distance is

computed between a vector produced for recognition and a vector previously produced for learning. The near proximity is acknowledged as a result of emotion recognition.

$$dist(x, y) = \sum_{i=1}^{n} |x_i - y_i| \quad (1)$$

5 Results

Results presented in this section are the outcomes of each model, and were then combined (in Section 5.2).

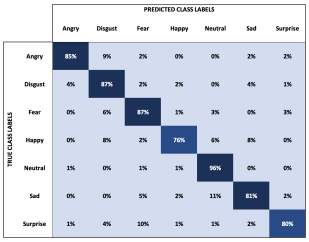
5.1 Unimodal Results

Outcomes presented in this section were achieved using above methods, and were achieved using both BAUM and SAVEE datasets. The latter was used to train the MER model, to place landmarks (better "in-the-wild" recognition of face landmarks was achieved using this dataset).

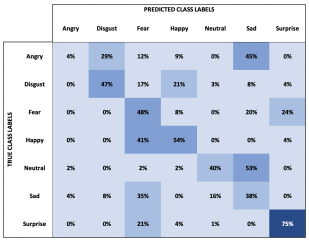
Modality	TER	SER	FER	MER
Accuracy	84.42%	43.71%	70.42%	68.71%

TABLE 2: Accuracy of each modality

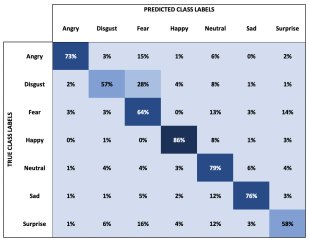
Also, confusion matrices from each model were plotted (Figures 9a to 9d) which detail in depth the obtained results.



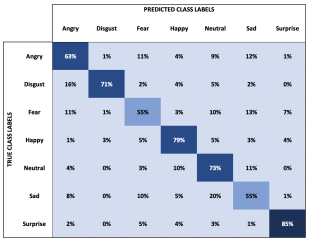
(a) Text Emotion Recognition Results



(b) Speech Emotion Recognition Results



(c) Face Emotion Recognition Results



(d) Motion Emotion Recognition Results

Fig. 9: Confusion Matrices

5.2 Multimodal Results

A weighted average was utilised to gather outcomes of the models, as shown on Equation 2. This basic average combines the accuracy for each feeling (for every stream), the recall of each emotion, for every model as well as the accuracy of the prediction, also for each deep learning method.

$$P_i = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} p_i \cdot mod_j \cdot rec_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} mod_j \cdot rec_{ij}} \quad (2)$$

where:

P_i = final prediction for emotion i,
M = number of distinct emotions (*fixed*),
N = number of types of data (*fixed*),
p_i = prediction for emotion i,
mod_i = accuracy of the model i (*fixed*),
rec_{ij} = recall of emotion i, for model j (*fixed*).

6 Discussion

Over the past ten years, multimodal approaches have been used in a number of scientific studies on emotion recognition. combining many modal signals, including physiological signs, audio and written language, facial and auditory features, and other combinations of these modalities. The deep learning algorithms used in this research directly contribute to a higher rate of deeper comprehension to enhance performance, accuracy, and confidence.

Global results presents a 74%-accuracy rate where some emotions are better understood than others.

A summary of the results in Figure 10 can aid in evaluating the model's performance on each specific emotion. The minimum accuracy, precision, and recall are 45.4%, 42.1% and 42.1% (for disgust) whereas happiness is the best recognised one with more than 80% for each result. Disgust and fear share face traits as well as surprise, which could explain their poor precision. Another reason could be the distribution of these specific emotions is the dataset used, as fear is present for only 4%, and surprise 5%. Also, most basic emotions (happiness, anger, sadness and neutral) were very well recognised, which might be because this emotion had the largest data coverage and because of their simple features (especially for FER and MER).

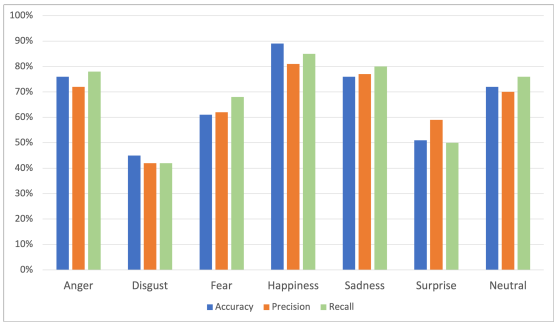


Fig. 10: Performance of the proposed multimodal approach for Ekman's basic emotions

Finally, as text and facial Emotion Recognition were the most accurate, such multimodal study could help to understand other mental states (e.g. sarcasm/irony if TER returns happiness and FER, anger/ sadness).

7 Conclusion

In this study, data from text, speech, image and motion capture was used to perform a multimodal emotion detection and recognition using BAUM and SAVEE datasets. The best individual structures for classification on each modality was achieved. Regarding the combination of results, a simple weighted average was integrated, but ensemble learning models or other methods could give in more robust and accurate results. A multimodal study has multiple benefits. First off, the lack of a modality would not impact the result as four modalities have been studied. Additionally, because of the optimisation of each modality (separately), the combined model takes a modular approach. This makes it possible to upgrade any individual model without affecting the other modalities in the combined model. Regarding the results obtained, state-of-the-art Text (DeBERTa) and Facial (ResNet-50) Emotion Recognition were achieved with satisfying outcomes. On the opposite, other models or modifications in their layers could benefit Speech (CNN) and Motion (3DCNN) Emotion Detection as other studies present better accuracy results.

Acknowledgment

The author would like to thank the University of Portsmouth, and specifically Dr. Alaa MOHASSEB and Dr. Ella HAIG for their support and useful advices.

References

[1] C. Baziotis, N. Athanasiou, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos, "Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning", *arXiv preprint arXiv:1804.06658*, 2018.

[2] "SemEval", https://competitions.codalab.org/competitions/17751, 2018.

[3] F. M. Alotaibi, "Classifying text-based emotions using logistic regression", 2019.

[4] S.-H. Park, B.-C. Bae, and Y.-G. Cheong, "Emotion recognition from text stories using an emotion embedding model", in *2020 IEEE international conference on big data and smart computing (BigComp)*, pp. 579–583, IEEE, 2020.

[5] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged svm over bert word embedding ensembles", *arXiv preprint arXiv:2010.01309*, 2020.

[6] J. Liscombe, J. Venditti, and J. B. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features", 2003.

[7] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on dnn-decision tree svm model", *Speech Communication*, vol. 115, pp. 29–37, 2019.

[8] S. Mao, D. Tao, G. Zhang, P. Ching, and T. Lee, "Revisiting hidden markov models for speech emotion recognition", in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6715–6719, IEEE, 2019.

[9] I. Shahin, A. B. Nassif, and S. Hamsa, "Emotion recognition using hybrid gaussian mixture model and deep neural network", *IEEE Access*, vol. 7, pp. 26777–26787, 2019.

[10] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection", in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5115–5119, IEEE, 2017.

[11] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition", in *Interspeech 2015*, 2015.

[12] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions", *arXiv preprint arXiv:1705.01842*, 2017.

[13] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression", in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pp. 94–101, IEEE, 2010.

[14] C. Pierre-Luc and C. Aaron, "Challenges in representation learning: Facial expression recognition challenges", 2013.

[15] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video", *arXiv preprint arXiv:1711.04598*, 2017.

[16] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition", *arXiv preprint arXiv:1802.06664*, 2018.

[17] X. Pan, G. Ying, G. Chen, H. Li, and W. Li, "A deep spatial and temporal aggregation framework for video-based facial expression recognition", *IEEE Access*, vol. 7, pp. 48807–48815, 2019.

[18] Y.-T. Lan, W. Liu, and B.-L. Lu, "Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism", in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, IEEE, 2020.

[19] H. Zhang, "Expression-eeg based collaborative multimodal emotion recognition using deep autoencoder", *IEEE Access*, vol. 8, pp. 164130–164143, 2020.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning", in *ICML*, 2011.

[21] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video", in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pp. 467–474, 2015.

[22] "Emotion Recognition in the Wild Challenge", https://sites.google.com/view/emotiw2020, 2020.

[23] J.-H. Lee, H.-J. Kim, and Y.-G. Cheong, "A multi-modal approach for emotion recognition of tv drama characters using image and text", in *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 420–424, IEEE, 2020.

[24] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique.", 1994.

[25] "AudioSegment from pydub library", https://audiosegment.readthedocs.io/en/latest/audiosegment.html, 2022.

[26] "Speech Recognition", https://pypi.org/project/SpeechRecognition/, 2022.

[27] "Video Capture", https://docs.opencv.org/3.4/d8/dfe/classcv_1_1VideoCapture.html#a57c0e81e83e60f36c83027dc2a188e80, 2022.

[28] "Noise Reduction from Audio file", hthttps://pypi.org/project/noisereduce/, 2022.

[29] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using mfcc", in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2257–2260, 2017.

[30] C. Busso, S. Lee, and S. S. Narayanan, "Using neutral speech models for emotional speech analysis.", in *Interspeech*, pp. 2225–2228, 2007.

[31] N. Hardeniya, *NLTK essentials - Text Preprocessing*. Packt Publishing, 2015.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", 2018.

[33] H. Liu, "Sentiment analysis of citations using word2vec", *CoRR*, vol. abs/1704.00177, 2017.

[34] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention", 2020.