International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 428-436, October-2023
https://doi.org/10.47001/IRJIET/2023.710057

EMOSENSE – Multi-Modal Emotion Recognition to Identify Emotions

[1]De Silva J.A.D.P.R, [2]Lanka P.A.C, [3]Jayawardena R.D.T.M, [4]Nandakumara K.S.S, [5]Lakmini Abeywardhana, [6]Dilshan De Silva

[1,2,3,4,5,6]Faculty of Computing (FoC), Sri Lanka Institute of Information Technology (SLIIT), Malabe, Sri Lanka
Authors Emails: [1]it20136642@my.sliit.lk, [2]it20137014@my.sliit.lk, [3]it20004354@my.sliit.lk, [4]it20135720@my.sliit.lk, [5]lakmini.d@sliit.lk, [6]dilshan.i@sliit.lk

***Abstract* - An extended study has been done over the past years to better comprehend human emotions. The embracement of technology to recognize and react to human emotions has become a required component of society. We present a fully functional multi-modal emotion recognition system in this study that integrates data from text, voice, facial expressions, and body language. In this study, the automatic classification of anger, fear, joy, sadness, surprise, disgust, and neutral emotions from text, facial expressions, voice, and body movements have been studied on the TESS, MELD, FER2013, and EDNLP datasets. Random Forest Classifier has been used for the classification of emotions using body language, VGG16 pre-trained model for facial emotion classification, Logistic Resgression for text emotion classification, and CNN for voice emotion classification. The logistic regression model for text emotion prediction leverages natural language processing (NLP) techniques to extract emotions from textual data. The CNN-based voice model utilizes speech recognition and emotion recognition algorithms to analyze audio signals and detect emotional cues in the speaker's voice. The facial expression model employs a combination of CNN-based VGG16 pre-trained model and modified convolutional layers to detect emotions. Meanwhile, the Random Forest Clasifier model is used to capture and interpret non-verbal cues, such as gestures, posture, and overall body movements, to enrich the emotion detection process. The real strength of our proposed system lies in its ability to synergistically combine information from multiple modalities.**

Keywords: multi-modal emotion detection, facial expressions, voice, text, body language, human emotions, machine learning.

I. INTRODUCTION

In the vast landscape of human and computer interaction, the objective of understanding and interpreting human emotions has grown very much. Emotions play a hugely important role in supporting human cognitive processes, influencing decision-making, and driving social interactions.

Emotions can be elaborated further by many simple statements. The psychological and physiological states known as emotions are intricate, and diverse, and occur in reaction to both internal and external stimuli. They have a fundamental impact on our thoughts, actions, and general well-being, contributing to the human experience.

Parallelly, the development of emotionally intelligent machines has become a main goal for researchers and technologies. Traditionally, the area of emotion recognition was solely based on relying on verbal and nonverbal cues, facial expressions, and other psychological responses. However, existing uni-model approaches suffer from accuracy problems. Which leads to failing to capture the complexity and nuances of human emotional states.

In recent years, the progress of multi-model approaches has created a huge impact on various fields. By leveraging multiple sources of information, such as facial expressions, speech patterns, body language, physiological signals, and even textual context, multi-modal systems strive to achieve a more comprehensive and contextually aware understanding of emotions.

The primary objective of this research paper is to modify and create a multi-model emotion recognition and present the potential of this application. Which will be discussed thoroughly throughout this paper.

II. RELATED RESEARCHES

There are several existing researches which contribute to a similar domain of our application. The technologies and approaches that have been used for each work have a significant difference.

When it is considered for emotion detection through facial expressions, there is existing research done on multiple face detection, high accuracy achievement, and usage of deep

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 428-436, October-2023

https://doi.org/10.47001/IRJIET/2023.710057

CNN. According to [1], Amit Pandey, Aman Gupta, and RadheyShyam have proposed a facial expression identification approach based on a Convolutional Neural Network (CNN) model that extracts facial features more effectively by the removal of the background of an image. They have used the FER-2013 dataset and some other datasets from multiple sources to increase the accuracy up to a percentage of 70%. When determining the number of layers for background removal CNN and face feature extraction CNN, they have assumed that the number of layers is directly proportional to accuracy and inversely proportional to execution time. Boddepalli Kiran Kumar, Korla Swaroopa, and Tarakeswara Rao Balaga have proposed a system to improve the process of facial sentiment analysis [2].

The proposed system uses five modules for its model design. They are; the face-capturing module, pre-processing module, training module, face recognition module, and expression recognition module. This research has accomplished good face detection and emotion extraction from facial photos. The suggested method in their research succeeded in using even low-resolution photographs and obtaining higher accuracy in output while being computationally efficient. The future improvements of this system according to [2] would be to use computer vision integrated for facial recognition and emotion identification and improve advanced feature extraction and classification in face expression recognition.

Emotion detection systems through voice are growing in popularity in the realm of the sentimental information technology industry, as they can enhance human-computer interaction by enabling machines to comprehend and react to human emotions. From this literature survey, we will compare several multi-modal emotion detection systems through voice with the other three components. In line with the study conducted by Jianhua and his colleagues [3], their setup utilizes deep neural networks (DNNs) to assess both auditory and visual aspects for identifying emotions. The auditory aspect is analyzed using Mel-frequency cepstral coefficients (MFCCs), while the visual aspect is analyzed using a convolutional neural network (CNN).

The findings indicate that the setup attains excellent precision in emotion identification. According to Gokul and teams' study [4], their arrangement utilizes feature-level merger techniques to combine elements extracted from multiple modalities, including voice, facial expressions, and physiological signals. The arrangement then utilizes a machine learning algorithm, such as a support vector machine (SVM), to categorize emotions. The outcomes show that the arrangement accomplishes great precision in emotion detection. In this research project, we are supposed to develop an emotion detection system to detect specific emotions using voice recognition integrated with other components which are facial, text and handwritten, and body movements.

In the fields of sentiment analysis and natural language processing (NLP), emotion identification from text is an active study subject. Researchers in this field have looked into a variety of approaches and methods. The followings are some important studies and related research on text emotion recognition: According to [5]created a new news aggregator to gather information from many news sources and classify the content's themes into eight emotion categories using semantic parsers and SenseNet[6].

[7] used the cognitive structure of emotional information to study natural language and affective information. To enhance communication in a text-based instant messaging system that uses emoticons or avatars that depict the detected emotion to describe the emotional state, they created the ALICE chatbot based on the AIML script. [8] used OMCS (Open Mind Common Sense), a database of 400,000 facts about ordinary life, to divide phrases into groups according to basic emotions (happy, sad, angry, afraid, disgusted, and startled). [9] created an emotion extraction engine that can examine the input language from a chat conversation, extract the emotion, and display the emotive image on the communicating users' screens. Their parser solely takes into account positive sentences, words without commencing auxiliary verbs (no inquiry sentences allowed), sentences in the present continuous tense, etc.

The development of machine learning models for emotion detection from body language is a rapidly evolving field, with significant potential for applications in areas such as human-computer interaction, healthcare, and security (Karg, Samadani, Gorbet, Kuhnlenz, Hoey, &Kulic, 2013) [10]. The model learns to identify patterns and correlations between body language and emotions, enabling it to predict emotions from new, unseen body language data. Feature extraction is a critical step in this process, where meaningful information is extracted from the raw body language data. This could include aspects such as posture, gesture speed, and movement fluidity (Kleinsmith & Bianchi-Berthouze, 2013) [11].

These features are then used as input for the machine-learning model. Various machine learning algorithms have been employed in this context, including Support Vector Machines (SVM), Decision Trees, and more recently, Deep Learning methods such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [12] (Kapoor, Qi, & Picard, 2003; Kahou et al., 2013) [13]. However, the challenge lies in the inherent subjectivity and complexity of

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 7, Issue 10, pp 428-436, October-2023
https://doi.org/10.47001/IRJIET/2023.710057

human emotions and their expression through body language. Emotions can be expressed differently by different individuals, and even by the same individual under different circumstances (Calvo &D'Mello, 2010) [14]. This variability makes it difficult to create a universally applicable model. Therefore, ongoing research is focused on improving the generalizability and robustness of these models, through methods such as transfer learning and data augmentation (Poria, Cambria, Bajpai, & Hussain, 2017) [15].

III. METHODOLOGY

3.1 Proposed Multi-Modal Emotion Recognition Architecture

3.1.1 Overall Model Architecture

The system consists of four components that combine facial expression analysis, body movements analysis, voice analysis, and handwriting text analysis to detect emotions in individuals. The facial expression analysis focuses on the analysis of facial muscle movements. Body movements analysis involves analysis of the gestures of the body to detect emotions. Voice analysis is the process of analyzing the tone, pitch, and other features of the voice of a person to determine the emotional state of the individual. Handwriting analysis involves the analysis of the text of a person's handwriting content to determine the emotional state of the person's handwriting.

3.2 Emotion Detection from Facial Expression

I. Image Acquisition and Pre-Processing

This component is trained based on two datasets that contain the same emotion labels. The two datasets are MELD and FER2013[16, 17] When considering the train split of the MELD dataset, each raw video is about 0.5 seconds. These videos are separated into 16 frames. Thereafter, the Viola-Jones algorithm is performed to detect faces in each frame. These faces are filtered to obtain the faces of the speakers by removing non-speaker faces. The detected faces are then labeled based on the CSV file of the MELD dataset and classified into separate folders based on emotion. The FER2013 dataset images are also added to the corresponding folders based on respective emotions. Thereafter, all these images are converted to grayscale and resized. These training image data are then further pre-processed using data augmentation. Thereafter, similar test data was prepared using both datasets.

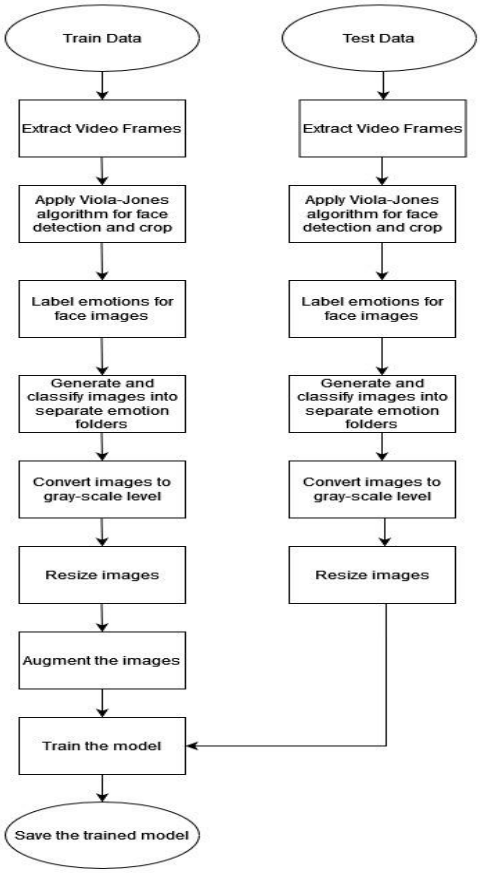

Figure 2: The structure of image acquisition and pre-processing

II. Proposed Architecture

The preprocessed images for this component are trained upon four models. They are; the pre-trained models of ResNet-50, VGG16, Inceptionv3, and a basic CNN model. The evaluation of each of these models will be discussed in the 'results and discussion' section.

According to [18] VGG16 model is a Convolutional Neural Network (CNN). It is a good vision model architecture that is used even at present for most research purposes. The model contains an arrangement of convolutional and pooling layers throughout the architecture and ends with two fully connected layers along with a softmax layer to generate the output.

In the proposed architecture, to ensure that only the convolutional base is used, the pre-trained VGG16 model is loaded without its fully connected layers. Thereafter, a series of new layers are added on top of the VGG16 layers. This includes a Flatten layer to turn the output of the convolutional base into a 1-dimensional vector. Then, two fully connected layers are added, each followed by Dropout layers to prevent overfitting, an Activation layer to introduce non-linearity with the ReLU function, Batch Normalization to increase training

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 428-436, October-2023

https://doi.org/10.47001/IRJIET/2023.710057

stability, and so forth. The final Dense layer with softmax activation is included to output probabilities for the classification of the image into one of the seven target classes. Thereafter, the merged model is prepared for training.

III. Emotion Detection in Real-Time

The facial expression emotion detection component is tested in real-time using a web camera. When using our implemented web application, the user could navigate to the corresponding web page which allows the user to detect their emotions through facial expressions. The user should allow the web camera access as a minimal requirement for this component to work. Here, the user could upload a 3-5 second video using the web camera in a way that shows their current facial expressions. Thereafter, the video will be split into frames and the trained model is used to predict the emotion in each frame. The final emotion of the video is taken as the average/overall emotion considering all frames. This output will be displayed to the user. This component allows the user to provide feedback on the emotional result provided by the application. Thereby it will lead to self-improvement of the training model. It is not mandatory to provide feedback because it will be a great disadvantage for people with alexithymia traits and people who have complete difficulty in identifying their emotions.

3.3 Emotion Detection from Body Language

I. Image Acquisition and Pre-Processing

This model is trained based on the MELD [17] dataset, which utilizes the MediaPipe Holistic landmarks detection algorithm to extract essential body landmarks, including pose and face landmarks, from video clips in the dataset. The MediaPipe Holistic model is a cutting-edge machine learning technique capable of detecting 468 face landmarks and 33 pose landmarks. The use of this comprehensive detection allows the capturing of a diverse set of physical cues associated with various emotions. The extracted landmark data includes X, Y, and Z coordinates, and visibility, which are organized into separate lists for pose landmarks and face landmarks. To facilitate further analysis, these lists are concatenated into a single row containing all the landmark coordinates for each frame. Subsequently, the landmark data is paired with its corresponding emotion label and exported into a CSV file. This pre-processing step is crucial in preparing the data for the subsequent machine learning model development. Thereafter, similar test data was prepared using the dataset.

II. Proposed Architecture

The preprocessed data for this component is trained using seven algorithms, namely LogisticRegression, RidgeClassifier, RandomForestClassifier, GradientBoostingClassifier, BaggingClassifier, AdaBoostClassifier, and ExtraTreesClassifier. The evaluation of each model will be discussed in the 'Results and Discussion' section.

III. Emotion Detection in Real-Time

The body language emotion detection component undergoes real-time testing using a web camera and recorded videos. Our web application enables users to access a dedicated webpage where they can detect their emotions through body language. Users have the option to either upload a video or use their web camera to capture their current body movements. Once the video is uploaded or captured, it is split into frames, and our trained model predicts the emotion for each frame. To determine the overall emotion for the entire video, we calculate the average emotion across all frames. This result is then displayed to the user.

Moreover, this component encourages users to provide feedback on the emotional results provided by the application. User feedback plays a crucial role in the continuous improvement of the training model. However, providing feedback is not mandatory, as it may present challenges for individuals with alexithymia traits or those who struggle with identifying their emotions.

3.4 Emotion Detection from Textual Method

I. Neattext-Based Pre-Processing

Kaggle Emotion Dataset for NLP[19] was used for training purposes. The dataset has eight emotion labels that are named as follows: joy, sadness, fear, anger, surprise, neutral, disgust, and shame. Before feeding the model, certain initial pre-processing tasks were completed. As mentioned earlier, Neattext is used to preprocess the dataset. Stopwords and other unwanted words will be removed from the text dataset by Neattext. Among the list of data-removing methods of Neattext, we used only the remove_stopwords and remove_userhandles. Logistic regression is used for both classification and regression. Because machines cannot process the row text data, Counter Vectorizer is used to convert the row text into matrix numbers. The train_test_split method is used during the splitting of the dataset. It splits the dataset into two sets, a train set, and a test set.

II. Proposed Architecture

The preprocessed dataset is trained using Naive Bayes, Random Forest, and LogisticRegression models. The ‘results and discussion' section will go over how each of these models fares when evaluated. According to Asghar et al. (2019) applied a variety of machine learning models on the ISEAR

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 428-436, October-2023

https://doi.org/10.47001/IRJIET/2023.710057

dataset, they discovered that the logistic regression model performed better than other classifiers [20]. The implemented logistic regression was a statical algorithm that was used for binary classification under supervised learning that predicts the probability of a binary outcome based on one or more independent variables. It gave an acceptable output when compared with the other models according to the changes made to specific hyperparameters. Logistic regression is named as the base model for this component. To reduce the model errors, and to do the model faster used a machine learning pipeline and it needed to specify the machine learning stages which are LogisticRegression and the CountVectoriizer to automate.

III. Performance of the Architecture to Detect the Emotion through text

This component automatically predicts the emotion by analyzing the user's submitted written things. The model is based on NLP. Used Scikit-learn, Pandas, and NumPy as the Python libraries to build the model. The model analyzes the text which is extracted from the image by using the OCR model which was created using OpenCV and the Pytesseract libraries. To predict the emotion, users should submit the image of the written things. Rather than that users can use the whiteboard space which is provided to write things. However, the user should take a screenshot of that and upload it. The component mainly focuses on the prediction of the emotion of the submitted written things by comparing the semantic textual similarity between the submitted text and the model text. This is highly helpful for those who exhibit symptoms of alexithymia. In this study, we investigated how to increase the accuracy of emotion recognition by combining text with additional modalities, such as photos, audio, and video. This method makes use of the extra context that various modalities offer.

3.5 Emotion Detection from Voice Expression

The voice-based emotion detection component aims to acknowledge and distinguish human sentiments by studying vocal hints and structures. Its primary goal is to provide users a way to perceive their emotional conditions through sound examination, alongside other approaches if relevant. This component offers an understanding of the precision, entirety, and general execution of the sound-based feeling identification system, empowering analysts to evaluate its efficiency and potential uses in different fields.

3.5.1 Enhancing the Architecture of Emotion Detection through Voice

I. CNN Based on Pre-Processing & Feature Extraction

The TESS dataset [21], greatly contributes to the CNN-based preprocessing method for emotion detection through voice. During the preprocessing technique, each voice clip from TESS is loaded and undergoes resampling to a fixed sample rate, ensuring uniformity throughout the database. For this component, preprocessed data are trained using four algorithms, specifically RNN, SVM, Random Forest Classifier, and CNN-based algorithm. By offering a diverse collection of voice recordings with precise sentiment labels, TESS acts as a valuable source of emotional speech data for training the CNN model. Additionally, normalization of the waveforms is performed to uphold a consistent amplitude range, which is vital for dependable feature extraction. The feature extraction step, where the sound waveforms are transformed into feature vectors, is improved by utilizing Mel-frequency cepstral coefficients (MFCCs) - a commonly used technique in speech analysis. These MFCCs capture the fundamental characteristics of speech, enabling the CNN model to effectively identify emotion-related patterns. Data augmentation techniques, if implemented, can further enhance the TESS dataset, improving model generalization and performance. After preprocessing, the dataset is divided into training, validation, and testing subsets to facilitate CNN model training and evaluation. Harnessing CNN-based preprocessing empowers the model to precisely detect emotion from voice data, making it a potent tool in real-world emotion recognition tasks.

II. Pre-training the model

To pre-train the CNN model for precise sentiment detection using the TESS dataset, initially, it uploads the audio data and implements CNN-based preprocessing methods containing resampling, normalization, and feature extraction using Mel-frequency Cepstral Coefficients (MFCCs). Then it divides the dataset into training, validation, and testing sets for training, hyperparameter tuning, and assessment, respectively. The model will be compiled with categorical cross-entropy loss and an optimizer such as Adam. Ultimately, it assesses the model on the testing data to evaluate its accuracy and performance. Optionally, it can conduct fine-tuning and hyperparameter tuning to further optimize the model's accuracy. So, the CNN model can be used to precisely recognize emotions from voice data using the TESS dataset for sentiment detection tasks.

III. Real-time performance of the Architecture

In this phase, we propose a multi-model approach that combines audio processing, machine learning, and real-time systems to achieve accurate and timely emotion recognition not only from voice inputs but also from video and text inputs. This system utilizes techniques such as feature extraction and

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 428-436, October-2023

https://doi.org/10.47001/IRJIET/2023.710057

model fusion to enable real-time processing of voice data. The proposed architecture consists of a front-end audio processing module for capturing and preprocessing voice signals, and multiple emotion recognition models trained on different datasets that means from our system the user can input their voice inputs to identify their current emotion. This is very useful to people who are suffering from alexithymia traits. This research contributes to the development of solutions for real-time emotion recognition through voice, with potential applications in varied components including human-computer interaction.

IV. RESULTS AND DISCUSSIONS

4.1 Emotion detection from facial expression

The four models that are evaluated for this component are ResNet50, VGG16, Inceptionv3, and a basic CNN model with 4 convolutional layers. The proposed model is a combination of VGG16 and modified convolutional layers. Table 1 shows the validation accuracy of each model conducted for 100 epochs corresponding to a learning rate of 0.01 and two different optimizers.

Table 1: Accuracy evaluation of models for two different optimizers

	Adam	SGD(lr=0.01, momentum=0.9)
VGG16	70%	68%
ResNet50	65%	68%
Inceptionv3	66%	67%
CNN model	62%	67%
Proposed model(Modified VGG16 model)	**70%**	**73%**

The changes made in the hyperparameter of the learning rate to 0.001 led to accuracy percentages as shown in table 2.

Table 2: Accuracy evaluation for a different learning rate

	Adam	SGD(lr=0.001, momentum=0.9)
VGG16	71%	73%
ResNet50	70%	72%
Inceptionv3	67%	70%
CNN model	66%	68%
Proposed model(Modified VGG16 model)	**74%**	**75%**

The model evaluations for accuracy show that the proposed model has more potential than other models to detect facial emotions. Some other evaluation matrices that are considered for these models are precision, recall, and f1-score. However, these evaluation matrices have also proved the improvement of performance in the proposed model. Accordingly, the modified VGG16 model (proposed model) shows a satisfactory output in most aspects when compared to other models. Thereby the modified VGG16 model is considered the base model for this component.

4.2 Emotion detection from body language

The seven algorithms that are evaluated for this component are LogisticRegression, RidgeClassifier, RandomForestClassifier, GradientBoostingClassifier, BaggingClassifier, AdaBoostClassifier, and ExtraTreesClassifier. The proposed model is RandomForestClassifier. Table 3 shows the validation accuracy of each model conducted.

Table 3: Accuracy and F1 score evaluations of models

	Accuracy %	F1 Score %
LogisticRegression	27.94	26.61
RidgeClassifier	37.69	35.88
GradientBoostingClassifier	56.53	55.90
BaggingClassifier	95.96	95.89
AdaBoostClassifier	28.13	27.23
ExtraTreesClassifier	96.06	95.97
RandomForestClassifier	**96.61**	**96.52**

The model evaluations for accuracy and f1 score show that the proposed model has more potential than other models to detect body language emotions.

RandomForestClassifier is known for its ability to handle complex relationships in the data, making it well-suited for tasks where the underlying patterns might not be easily captured by simpler models like LogisticRegression. Its ensemble nature, which combines multiple decision trees, allows it to capture both linear and non-linear relationships between features and the target variable.

Emotion detection from body language could be prone to noise due to variations in human expressions, environmental factors, and subjective interpretations. RandomForestClassifier's aggregation of multiple decision trees allows it to be robust to noisy data. Outliers or inconsistencies in the training data are less likely to impact the overall performance of the model.

In conclusion, the RandomForestClassifier's strong performance in emotion detection from body language might stem from its ability to handle complex patterns, robustness to noisy data, and the ensemble nature that helps mitigate overfitting. However, it's essential to consider other factors like interpretability, computational efficiency, and domain-specific requirements when choosing a final model. Further experimentation and analysis could provide deeper insights into the specific features and patterns that contribute to the RandomForestClassifier's success in this particular task.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 428-436, October-2023

https://doi.org/10.47001/IRJIET/2023.710057

4.3 Result of the Text Classification Model

For this component, Native Bayes, Random Forest, and Logistic Regression are the three models that are tested. Logistic regression is the foundation of the suggested model. The validation accuracy of each model, measured over 100 epochs, is displayed in Table 4.

Table 4: F1 Score of the trained models

Approach	Accuracy	F1 Score
Naive Bayes	67.02%	0.6702
Random Forest	63.72%	0.6372
Logistic Regression	69.35%	0.6935

The model that performed the best for text-based emotion identification was LogisticRegression. The model successfully extracted emotional cues from text with an accuracy of 69% in emotion categorization. The model's high accuracy reveals both how well it was trained and how well it might be able to recognize emotions. By contrasting the predicted emotion labels with the ground truth labels from the test dataset, the accuracy of the model was determined.

Table 5: Overview of the accuracy and training time

Dataset	Accuracy %	Training Time
EDNLP (Emotion Dataset For NLP)	69	~100s

The text from the photos was successfully retrieved by the OCR model. The system's capabilities were increased by this integration, which allowed it to interpret text that was retrieved from images that expressed emotions. The accuracy and effectiveness of the OCR module's text extraction process influence how well the emotion detection system works. The trained model outperformed or was on par with existing research in its classification of emotions, with an accuracy of 69%. Zhang et al. (2018) used CNNs and other deep-learning techniques to increase the accuracy of emotion categorization by taking contextual data into account [22].

The test findings showed the system's potential for accurately recognizing emotions from a variety of sources, adding to the field of emotion detection and its practical applications. However, when considering computational time with performance, LogisticRegression is the most suitable model for the scenario.

4.4 Emotion detection from voice expression

When contrasting the CNN model to other models for emotion detection through voice, CNNs have demonstrated significant advantages. CNN model is very useful to surpass in learning hierarchical characteristics from spectrogram or MFCC portrayals of audio data with catching crucial patterns for sentiment recognition. Their capacity to handle spatial information in the data makes them well-suited for speech-related tasks. RNN model, while capable of capturing temporal connections, may encounter difficulties with long-range connections in audio data. While trying with using SVM models, though interpretable, might not achieve the same precision as CNNs due to their reliance on handcrafted characteristics and restricted capacity for complex datasets. Other than the above models, Random Forest models can be effective for specific issues but may fall short when dealing with unprocessed audio data. CNNs, on the other hand, perform well when trained on extensive and diverse datasets, enabling them to accomplish good precision in sentiment identification. Its capacity to automatically learn characteristics from the data, combined with its scalability, makes them more convenient and powerful for accurately detecting sentiments through voice data.

TESS dataset, adapting the learned representations to the target domain. The CNN model can benefit from the pre-learned features, improving its ability to detect emotions accurately in both audio and multimodal data from the TESS dataset. The below table provides an overview of the accuracy and training time of emotion detection through voice and the models trained on the TESS dataset over multiple epochs.

Table 6: An overview of the accuracy and training time

Dataset	No. of Epochs	Model	Accuracy %	Training Time
TESS	50	RNN	~72	~1.5hrs
		SVM	~68.9	~0.5hrs
		Random Forest	~65.5	~1hr
		CNN	**~81.7**	**~2hrs**

V. CONCLUSION

This emotion identification modal specializes in emotion detection from diverse modalities, such as facial expressions, body language, textual data, and voice expressions. For facial emotion recognition, a proposed model combining VGG16 and extra layers outperforms different models, attaining the highest accuracy. In body language emotion detection, RandomForestClassifier proves to be the handiest model. The text-based total emotion detection makes use of LogisticRegression and achieves competitive accuracy. For voice expression emotion detection, CNN shows the best advantages over other models and attains appropriate accuracy whilst skilled at the TESS and MELD datasets. Overall, those architectures offer promising consequences in emotional reputation, with capacity programs in numerous domains like human-computer interplay and sentiment analysis. As for future works, we suppose to create an accurate multi-modal

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 428-436, October-2023
https://doi.org/10.47001/IRJIET/2023.710057

emotion identifier that involves integrating output formats across facial expression, body language, text analysis, and voice expression components.

REFERENCES

[1] A. G. R. S. Amit Pandey, "FACIAL EMOTION DETECTION AND RECOGNITION," *International Journal of Engineering Applied Sciences and Technology*, vol. 7, no. 1, pp. 176-179, 2022.

[2] K. S. R. B. Boddepalli Kiran Kumar, "Facial Emotion Recognition and Detection Using CNN," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 14, pp. 5960-5968, 2021.

[3] Z. Y. P. C. S. N. Jianhua Zhang, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," 2020.

[4] N. C. K. P. N. B. J. A. Gokul Subramanian, "Multimodal Emotion Recognition Using Different Fusion Techniques," 2021.

[5] M. Shaikh, H. Prendinger, and M. Ishizuka, Emotion sensitive.

[6] M. Shaikh, H. Prendinger, and M. Ishizuka, SenseNet: A linguistic.

[7] M. Shaikh, H. Prendinger, and M. Ishizuka, A cognitively based.

[8] H. Liu, H. Lieberman, and T. Selker, A model of textual affect.

[9] A. C. Boucouvalas and X. Zhe, Text-to-Emotion. Engine for Real.

[10] M. S. A. A. G. R. K. K. H. J. &. K. D. Karg, "Body movements for affective expression: A survey of automatic recognition and generation.," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 341 - 359, 2013.

[11] A. &. B.-B. N. Kleinsmith, "Affective body expression perception and recognition: A survey.," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15-33, 2013.

[12] A. Q. Y. &. P. R. W. Kapoor, "Fully automatic upper facial action recognition.," *In Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 195-202, 2003.

[13] S. E. B. X. L. P. G. C. M. V. K. K. .. &. B. Y. Kahou, "Emonets: Multimodal deep learning approaches for emotion recognition in video.," *Journal on Multimodal User Interfaces*, vol. 7, no. 1, pp. 99-111, 2013.

[14] R. A. &. D. S. Calvo, "Affect detection: An interdisciplinary review of models, methods, and their applications.," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18-37, 2010.

[15] S. C. E. B. R. &. H. A. Poria, "A review of affective computing: From unimodal analysis to multimodal fusion.," *Information Fusion*, vol. 37, pp. 98-125, 2017.

[16] C. B. T. C. Amil Khanzada, "Facial Expression Recognition with Deep Learning".

[17] D. H. M. N. C. M. Soujanya Poria, "MELD:A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," vol. 2019, June.

[18] D. J. RENJU RENJITH, "Emotion detection using facial expression recognition based on VGG16 network," *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 8, no. 7, pp. b934-b938, 2021.

[19] https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp.

[20] Asghar MZ, Subhan F, Imran M, Kundi FM, Shamshirband S, Mosavi A, Csiba P, Várkonyi-Kóczy AR (2019) Performance evaluation of supervised machine learning techniques for efficient detection of emotions from online content. arXiv preprint arXiv:190801587.

[21] "Toronto emotional speech set (TESS)," [Online]. Available: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess.

[22] Zhang, Y., Ishibuchi, H., and Wang, S. (2018). Deep Takagi–sugeno–kang fuzzy classifier with shared linguistic fuzzy rules. IEEE Trans. Fuzzy Syst. 26, 1535–1549. doi: 10.1109/TFUZZ.2017.2729507.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 428-436, October-2023

https://doi.org/10.47001/IRJIET/2023.710057

Citation of this Article:

De Silva J.A.D.P.R, Lanka P.A.C, Jayawardena R.D.T.M, Nandakumara K.S.S, Lakmini Abeywardhana, Dilshan De Silva, "EMOSENSE – Multi-Modal Emotion Recognition to Identify Emotions" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 10, pp 428-436, October 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.710057

****

