Positive Vibes: A Real-time Facial Emotion Detector And Output Based Task Recommender

RutujaTambe[1], Anmol Thete[2], Aniket Gadhave[3], Sumedh Dhengre[4]

[1,2,3]Student, Dept. of Computer Engineering, AISSMS College of Engineering, Maharshtra, India.
[4]Assistant Professor, Dept. of Computer Engineering, AISSMS College of Engineering, Maharshtra, India.

---***---

Abstract - *Emotions are an essential element of human interaction and communication. Although emotions are known to everyone, it is difficult to describe feelings. The Greek philosopher Aristotle thought of the emotion as a stimulus that explores the experience of gain or pleasure. Artificial Intelligence has been witnessing the tremendous growth in closing the gap between human and mechanical skills. The Computer Vision agenda empowers machines to view the world the way people do, see it in the same way and apply this information for many tasks such as image and video recognition, image analysis and segmentation, Media Recreation, Recommendation Programs, Natural Language Processing (NLP), etc. Emotional Awareness draws your attention on key research to solve many problems. However, there is research done on the detection of emotions and ways to experience stress, but not on ways to reduce stress technically. Therefore, the purpose of this paper is to create a system that detects a person's emotions through facial expressions using Deep Learning techniques such as Convolutional Neural Networks and to provide appropriate recommendations based on Emotions. Using real-time facial expressions, a number of emotions will be detected such as sadness, happiness, anger, fear, surprise and neutrality. Detailed facial expressions will be captured, and appropriate emotions will get detected using the CNN algorithm. Emotional analysis by CNN will help inculcate positive vibes among people by proposing mood enhancing activities.*

Key Words: **Convolutional Neural Networks, Deep Learning, Facial Emotion Recognition, real-time, Human-Computer Interaction, Haar feature-based Cascade Classifier.**

1. INTRODUCTION

Depression today is often portrayed as a major health concern. Depression should be diagnosed before it becomes a major health problem. According to the World Health Organization, stress is a major problem of our time and affects people's physical and mental health. Automatic human recognition has received a lot of attention recently with the introduction of IoT and smart environments in hospitals, smart homes and smart cities [7]. Machine learning algorithms have changed several fields on the computer and beyond, including Human Computer Interaction (HCI). If we look at the automatic detection system, however, the available systems are inadequate and do not have the ability to accurately identify emotions.

Facial expression is an important means of communication between people [6]. Therefore, facial expressions are considered to be the main means of obtaining emotions. High-resolution facial recognition is done in real-time applications for many applications, such as behavioral analysis, machine vision and video playback [7]. If computers can detect these emotional inputs, they can provide clear and appropriate assistance to users in ways that are most relevant to users' needs and preferences [16]. Emotions also alter the state of the human brain, and can directly or indirectly influence a number of processes.

The emotional impact on the appearance of the face or speech can be easily suppressed, and the emotional state is reflected in the functioning of the nervous system. It is acknowledged that emotional ability is an important factor in the next generation of personal robots, such as the Sony AIBO [22]. It has the ability to play a significant role in "smart rooms" and "active computer trainer". Some theoretical analysis uses icons as labels to reduce reliance on emotional classification in computer-aided teaching techniques [1]. CNN classification is often used for machine learning because it is easy to use. The CNC used five tweets, Naive Bayes, Support Vector Machines (SVM), neural networks, random forest, and pressure separation and pressure detection [1].

This paper is organized as follows: -

Section II involves the existing research conducted in the field of emotion recognition, whereas Section III includes the design and implementation of the proposed system. Section IV shows the experimental results obtained.

2. LITERATURE SURVEY

The social media data like posting on Facebook profiles, twits on twitter, etc. are used to recognize human stress, as people communicate their social media feelings, promote the acquirement of social data, and detect stress based on their behavior [1]. For stress recognition processes, Textual, visual and social attributes are measured accordingly. In the neural psychological stress sensing method [2], the author uses real details about the online micro-blog to identify relations between stress and users' tweet contents.

This also sets out two types of stress-based attributes Simple text, photos, social, and user stats are based on low content features from the micro blog's weekly post. The goal is to detect the complex events in uncomplicated videos of the Internet in [3]. The software proposed in [4] uses the Gaussian Method to categorize the stress of a young person based on a variety of functions taken from the tweets. The program is designed to provide parents with data about the stress levels of their children through mobile text messages. It also provides inspirational stories to understand troubled children and increasing stress.

In the implementation of the program in [5], the Stanford SNAP database is used, which yields widespread mood performance. The best result achieved so far is by passing the training images through Histogram of Oriented Gradients (HOG), followed by characterization by SVM, which gives an average precision of 85% [6]. The work in [7] aims to classify physically disabled people (deaf, dumb, and bedridden) and Autism children's emotional expressions based on facial landmarks and electroencephalograph (EEG) signals using a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) classifiers by developing an algorithm for real-time emotion recognition using virtual markers through an optical flow algorithm that works effectively in uneven lightning, different backgrounds, and various skin tones.

There have been efforts to develop Multi-modal information using facial expressions [8]. Haar-like features that efficiently use the Ada Boost cascade classifier, to detect any object in a given image or a video sequence involving human faces can process 384 x 288 pixels of a face image in approximately 0.067 seconds [9]. Some of the classification algorithms like K-Nearest Neighbour, Random Forest are applied in [10] to classify emotions. Deep RNN like LSTM, Bi-directional LSTM modeled for audio visual features are used in [11].

Facial emotion recognition can also be performed using image spectrograms with deep convolutional networks which is implemented in [12]. Real-time facial emotion recognition is done through RGB image classification using transfer learning methodologies in which knowledge gained from solving one problem is used for the implementation of another problem [13]. Emotion has been recognized from facial expressions using hidden Markov models and deep belief networks with Unweighted Average Recall (UAR) of about 56.36% approximately [14].

Inception Net is used for expression recognition with Kaggle (Facial Expression Recognition Challenge) and Karolinska Directed Emotional Faces datasets in [15]. Reference [16] analyzes the strengths and the limitations of systems based only on facial expressions or acoustic information. It also states two approaches which have been used to fuse these two modalities, which are decision level and feature level integration. Based on psychological studies [17], which show that visual information modifies the perception of speech, it is possible to assume that human emotion perception follows a similar trend.

Motivated by these clues, De Silva et al. conducted a few experiments [18], in which 18 people were a requisite to recognize emotion using visual and acoustic information separately from an audio-visual database recorded from two subjects. Several approaches to recognize emotions from speech have been reported, and a comprehensive review of these approaches can be found in [19] and [20]. Using short-spoken sentences, the authors of [21] recognized two kinds of emotions: approval or disapproval.

They conducted several experiments with features extracted from measures of pitch and energy, obtaining an accuracy ranging from 65-88 %. The main limitation of the global-level acoustic features is that, they cannot describe the dynamic variation along an utterance. According to [22], there is no previous study that has demonstrated a physiological signal-based emotion recognition system that is applicable to multiple users. Researchers at IBM recently reported an emotion recognition device based on mouse-type hardware [23].

3. PROPOSED SYSTEM

3.1. System Architecture

In this section, a step-by-step approach towards the various processes taking place for fulfilling this work has been described as per Fig. 1. The final prediction of emotion is preceded by multiple processes.

i. Face detection and Feature Extraction

The first step will be to determine the face of the person in the given input image, which is then succeeded by identifying the features (eyes and mouth). These features are then passed through their respective filters and transformation.

ii. Classification and Emotion Detection

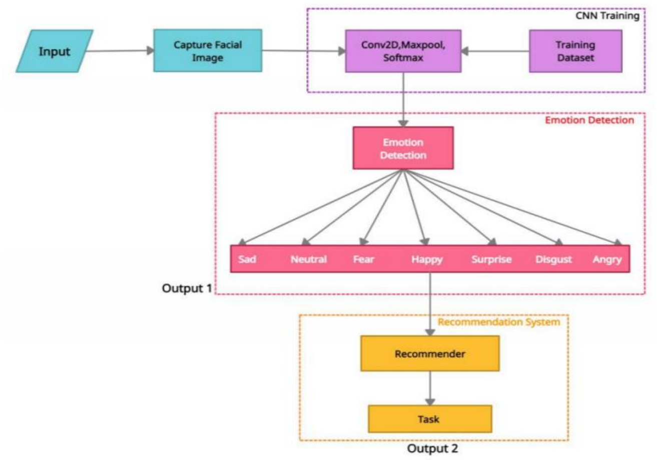


Fig -1: System Architecture

The outputs generated from the filters are then sent to the classifiers to get classified according to the trained data. This gives us the emotion predicted by the system.

iii. Task Recommendation

The user will be prompted to select the detected emotion from the given list of 7 emotions. Based on the selected emotion, the user will receive relevant suggestions.

3.2. Experimental Setup

The aim of this project was to detect the most appropriate human emotion through facial expressions as output, by providing real-time images of a person's face as an input. Also recommend a relevant task or song to enhance the person's mood and relieve them from stressful situations. Fig. 2 shows the complete flow of the system. The face of a person is captured, and different features are detected. The OpenCV python library was used for image processing and the human face was detected through Haar feature-based cascade classifiers.

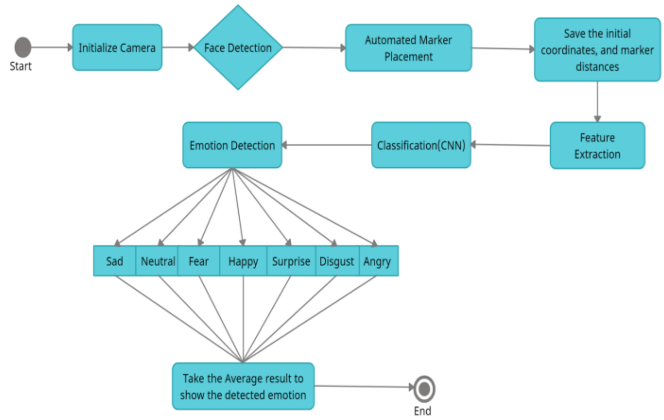


Fig -2: State Diagram

Our system was trained with dataset "fer2013.csv", and with every epoch the accuracy of the system increases. Then the model is serialized to JSON and the model weights are saved in an h5 file so that we can further make use of this file to make predictions rather than training the network again. Real-time images are taken on mobile phones or any other device, and then processed through the system accordingly. The interaction between the user and the application will be retrieved in the database. This data is used in implementing collaborative filtering using CNN algorithm. The result is shown to the user through reliable user interface.

3.3. System Requirements

Windows 7/8/10 or Linux Operating system can be used for developing the system. Python coding language (version 3.8.2) was used for programming the module. PyCharm IDE was used for development, training and testing. The TensorFlow API was used for building and deploying the model. Numpy, Pandas, Keras, OpenCV, are the various Python libraries used for different purposes. Keras is used for model fitting, developing the deep neural networks and OpenCV for image analysis and facial recognition. Tkinter (Python GUI Library) was used to develop the user interface for the system.

3.4. Algorithm

The Convolutional Neural Network (CNN) is a Deep Learning algorithm that can capture images, assign significance value (readable and discriminatory metrics) to the various aspects or elements in the image and have the capability of discriminating one from the other. The pre-processing required in the Convolutional Network is very trivial as compared to any other classification algorithm. While the previous filtering techniques are hand-engineered, with adequate training, Convolutional Networks have the ability to study these features. The architecture creates a better balance in the database due to the decrease in the number of parameters involved and the reuse of weights.

Hence, we can conclude that, the network can be trained to understand image perception in a much better manner. In cases of extremely basic binary images, the feed-forward neural networks might show an average precision score while performing prediction of classes, but would show very little to no accuracy when complex images having pixel dependencies throughout, comes into the picture. CNN is however able to successfully capture Spatial and Temporary dependencies on an image using the appropriate filters.

Different layers of the CNN architecture are as follows:

i. Input

We have an RGB image which has been captured by the user using the device camera. This image will then be converted to grayscale for simplification of the classification process. The major role of ConvNet includes reducing images into an easy-to-process form, without losing those of the most important factors in obtaining good forecasts.

ii. Convolutional Layer

The element which carries out the convolution operation in the first part of a Convolutional Layer is termed as the Kernel/Filter. Each filter will have a minimum height, width and depth which is equal to the input volume (3, if the input layer is an image). The possible dimensions of the filter can be a*a*3, while the 'a' can be 3, 5, 7, etc, but relatively smaller in size. The filter moves right through a certain number of Strides until it conveys the full width. Moving forward, it moves to the beginning (left) of the image with the same Stride Number and repeats the process until the whole image is traversed. The purpose of Convolution Operation is to extract high-quality features such as edges, in the input image.

iii. Pooling Layer

This layer is responsible for shrinking the spatial size of the Convolved Feature. This minimizes the computational power needed for processing data by dimensionality reduction. As shown in Fig. 3, pooling consists of two types: Max Pooling and Average Pooling. Max Pooling returns the highest value on the part of the image covered by the Kernel. Average Pooling, on the other hand, returns an average of all values from the image component compiled by Kernel. The Convolutional and Pooling layers combined together, form the i-th layer of the Convolutional Neural Network. Depending on the complexity of the images, the number of such layers may be increased to obtain very low level data, but it costs more computational power. Max Pooling methodology has been implemented in our system.

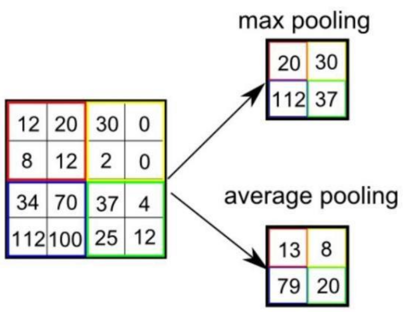


Fig -3: Types of Pooling

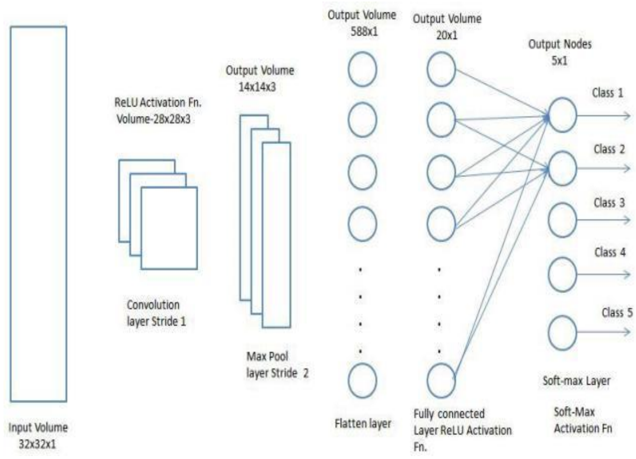


Fig - 4: Fully Connected Layer

iv. ReLu Activation Layer

In the ReLu (Rectified Linear Unit) Activation Layer, the corrective function is used to increase non-linearity on CNN. The data set is made up of different objects which are not linear to one another. Under this function, the grouping of information can be seen as a linear problem, although it is a non-straight problem.

v. Fully Connected Layer

Adding a fully connected layer is often a cost-effective way to learn the non-linear combinations of high-level features as represented by the output of Convolutional Layer Structure. After converting the input image into a suitable form for Multi-Level Perceptron, as depicted by Fig. 4, we flatten the image into a column vector. The flattened output is fed to a feed-forward neural network and back-propagation is applied to every iteration of training. Over a series of epochs, the model will be able to discriminate between dominating and certain low-level features in images and classify them using the Softmax Classification technique.

3.5. Implementation

Step 1: Collection of a dataset of images

In this project, we have used 'FER2013' database which consists of 35887 pre-cropped, 48*48 pixel grayscale images of faces, each labeled with one of the 7 emotion classes viz. anger, disgust, fear, happiness, sadness, surprise, and neutral.

Step 2: Pre-processing of images

The dataset contains two columns, "emotion" and "pixels". The "emotion" column includes numerical codes ranging from 0 to 6, inclusive, which represents the emotion present in the image. The "pixels" column contains a string enclosed by quotes for each image. The contents of this string include space-separated pixel values in row major order.

Step 3: The cropped face is converted into grayscale images.

Step 4: The pipeline ensures that every image can be fed into the input layer as a (1, 48, 48) Numpy array.

Step 5: The numpy array gets passed into the Convolution 2D layer.

Step 6: Pooling method called MaxPooling2D that uses a (2, 2) windows across the feature, map only the maximum pixel value. The Softmax function represents itself as a probability value for each emotion class.

3.6. Deployment

The system has been deployed as a web application, which takes the real-time image of user as the input, passes them to the model, retrieves the results and sends them back as a response to the user. We have used Keras with Tensorflow as back-end for building Neural Networks.

The layers added are as follows :

1. Convolution layer

This layer converts all pixels in its reception field into a single value. The convolution applied to the captured image, will reduce the image size and bring all the information in the field into one pixel. The end result of a convolutional layer is a vector.

2. Pooling layer

This layer is used to reduce the size of the feature maps. Therefore, it reduces the number of learning parameters and the amount of computation performed in the network. The pooling layer summarizes the features present in the feature map region formed by the convolution layer.

3. Batch normalization

It is a method of training very deep neural networks that standardize the inputs which get embedded in each mini-batch layer. This has the effect of stabilizing the learning process and dramatically reducing the amount of training time required to train deeper networks. It applies a change that keeps the mean output closer to 0 and the standard deviation output is closer to 1.

4. Activation Layer

The neural network activation function describes how the input weight is converted from a single node or a number of nodes in the network layer. It allows models to learn faster and perform better.

5. Flatten Layer

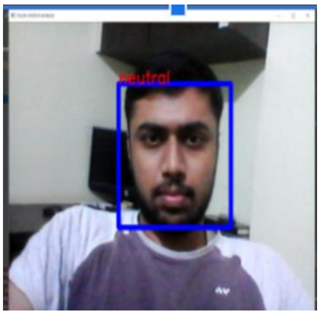
Flattening is the process of converting data into a 1-dimensional array, to be transferred as input to the next layer. We perform flattening of the convolutional layers to create a single vector of a long feature and connect it to a final classification model, called a fully connected layer.

The model works on a server as an API using TensorFlow, that handles all incoming emotion detection requests. It then transfers those requests to a trained model that resides on a web server. The model processes the input from the application and returns the emotion that is detected. The user then needs to select the emotion which has been detected by the system from the given options, and then they will receive the recommended tasks for that particular emotion.

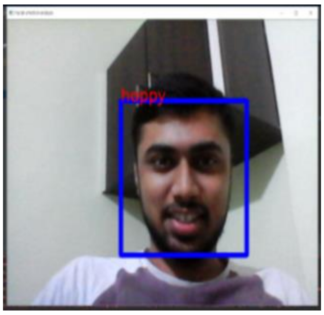
4. RESULTS

Emotions are an integral way of expressing our judgments and decisions in daily life, and this work aims to identify and directly detect these feelings. Although the use of markers is not suitable for actual use, the analysis presented in this paper provides important indications about the emotional discrimination contained in the various face blocks. The results presented in this paper show that areas such as the eyes and mouth, provide a good sense of emotional classification such as neutrality, happiness, fear and sadness. While, eyebrows play an important role in other emotions such as surprise, anger and disgust. Fig. 5 represents the 7 different emotions detected by the system.

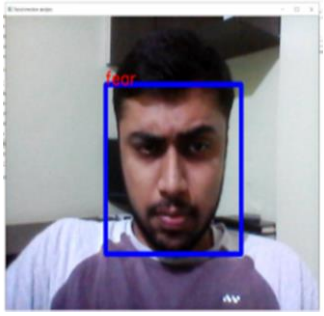
As seen in Fig. 6, the system provides some relevant suggestions to the users, as per the emotion detected, in order to enhance their mood. For detecting emotions using facial expressions from the input image, the “fer2013.csv” data set was used to train the model. The model was first run upto 50 epochs, where an accuracy of 75.53% was achieved. Later on, it continued to run until 100 epochs to obtain a higher precision value, and the accuracy improved to 84.99%. Emotion detection using facial expressions has a direct relationship with the number of epochs. Therefore, the system is able to detect emotions with the accuracy of “84.99%” using the proposed model in 100 Epochs. The beauty of this system is that, the image uses only 2 important facial features, namely the eyes and mouth, to get the perfect facial expression, which adequately reduces the amount of final data needed for future testing and use. Based on a study of various emotion recognition systems, as well as comparisons with the system suggested in this paper, we come to the conclusion that emotional detection through emotional recognition using real-time human images, is an excellent uni-modal system for the intentional action.



Emotion 1: Neutral

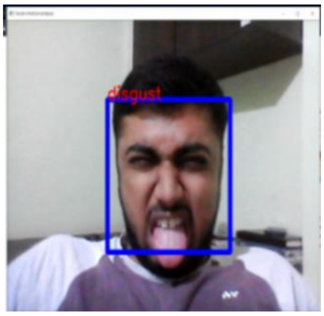


Emotion 2: Happy

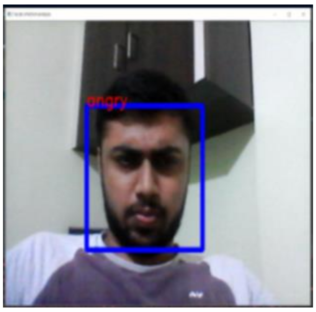


Emotion 3: Fear

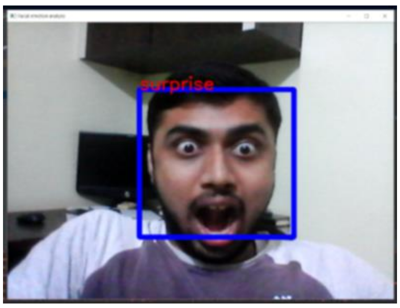
Emotion 4: Sad



Emotion 5: Disgust

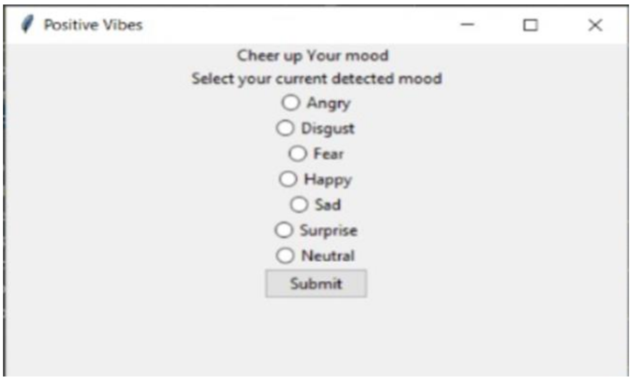


Emotion 6: Angry

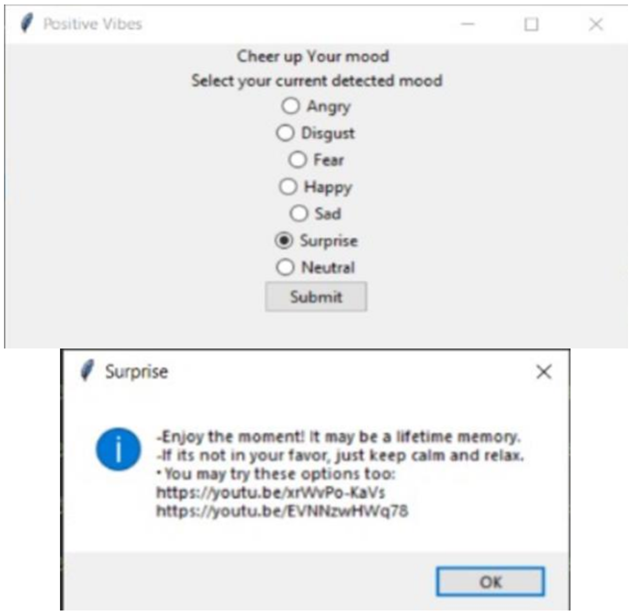


Emotion 7: Surprise

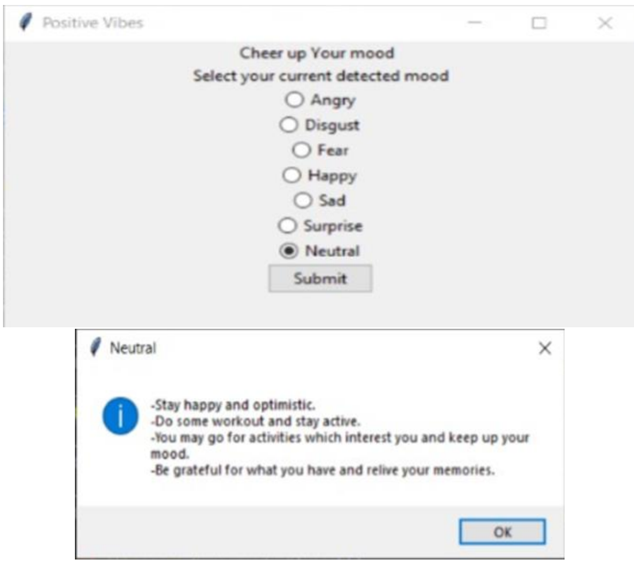
Fig. 5: Emotions Detected



User prompt for selecting the detected mood



Suggestions for 'Surprise' emotion



Suggestions for 'Neutral' emotion

Fig. 6: Recommended Tasks

5. APPLICATIONS

'*Positive Vibes*' is a user friendly application. Users can get instant feedback and customized recommendations using the internet or mobile phone easily. It does not require any additional hardware. This application may help employees and the HR (Human Resource) team of any company to manage stress levels. This will create a healthier work environment and increase productivity. Employees and managers can also restore the positive and negative feelings of employees and customers who help businesses grow. Adjusting the difficulty level in a game with prediction to improve attraction and business value, can be made possible using 'Positive Vibes'. Automatic Detection of Cheating during online tests or exams, using posture and Emotional Analysis can prove to be a useful application of the proposed system. It can also be utilized for capturing students' attention in the e-learning environment.

6. CONCLUSIONS

Facial detection is considered to be one of the most complex problems in computer vision, due to the large differences caused by changes in lighting, appearance and speech [6]. This study analyzed the strengths and weaknesses of emotion detection through the analysis of social media using data sets and acoustic emotion classifiers. In such kind of systems, some pairs of emotions are often classified incorrectly. A model named 'Positive Vibes' has been suggested to detect the facial emotion of people, which will then provide suitable recommendations based on the emotion detected. The "fer2013.csv" dataset is used as training content and CNN is used for classification.

The proposed work is capable of recognizing 7 integral emotions – Happy, Sad, Anger, Fear, Neutral, Disgust and Surprise; with the help of the Convolutional Neural Networks algorithm showing an accuracy of 84.99% when run for 100 epochs. This model can drive business outcomes and judge the emotional responses of the audience. Based on the research conducted in the presentation of this paper and the results obtained through the proposed system, it is possible to view greater accuracy using audio and visual methods. Therefore, the next generation of human-computer interface can detect human responses, and respond appropriately and efficiently to changing user contexts, improving the performance and engagement of current social networks.

7. FUTURE SCOPE

There was a lot of research and studies conducted on emotion recognition, where deep learning methods were used to identify emotions. It is a future requirement to have a more reliable model, with unlimited possibilities across all sectors. For future work, the accuracy of the proposed system and precision can be improved by collecting large amounts of data from multiple studies. The next steps in this study will be to find better ways to fuse visual and acoustic information that will transform the power of facial expressions along with speech.

Acoustic information in the form of segments can be used to detect emotions at an independent frame level. Also, it may be helpful to find another type of factor that describes the relationship between the two approaches in terms of temporal progression. For example, the relationship between movement of facial muscles and the pitch contour along with the energy can assist in discriminating emotions. Finding emotions in real-time videos, will also be part of the ongoing research.

ACKNOWLEDGEMENT

The authors would like to acknowledge the partial support of the project guide *Prof. S. G. Dhengre,* serving as an Assistant Professor in the All India Shri Shivaji Memorial Society's College Of Engineering, Pune-01. His continual support and guidance, along with beneficial suggestions stood noteworthy in the successful completion of this paper.

REFERENCES

[1] S. Meshram, Prof. R. Babu , Prof. J. Adhikari, "Detecting Psychological Stress using Machine Learning over Social Media Interaction", in Proceedings of the Fifth International Conference on Communication and Electronics Systems, 2020.

[2] H. Lin, J. Jia, J. Qiu, Y. Zhang, L. Xie, J. Tang, L. Feng, T. Chua, "Detecting Stress Based on Social Interactions in Social Networks", IEEE Transactions on Knowledge and Data Engineering, 2017.

[3] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, Y. Yu, "Semantic concept discovery for large-scale zero-shot event detection", in Proceedings of International Joint Conference on Artificial Intelligence, 2015.

[4] Q. Li, Y. Xue, J. Jia, and L. Feng, "Helping teenagers relieve psychological pressures: A micro-blog based system," in EDBT, 2014.

[5] A. Mogadala, "Twitter User Behavior Understanding with Mood Transition Prediction", 2020.

[6] C. Jain, K. Sawant, M. Rehman, R. Kumar, "Emotion Detection and Characterization using Facial Features", in 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering, November 2018.

[7] A. Hassouneh, A. M. Mutawa, M. Murugappan, "Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods", Published by Elsevier Ltd. in June 2020.

[8] P. Ekman, "Darwin and facial expression: a century of research in review", United States Of America: Academic Press Ishk; 2006.

[9] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", In International Conference on Image Processing (ICIP), 2001.

[10] A. M. Yusuf, M. B. Mustaafa, M. Malekzedeh, "Emotion recognition research: an analysis of research focus," Int. J. Speech Technol., vol. 0, 2018.

[11] L. Cavidon, H. M. Fayak, M. Lich, "Evaluating deep learning architectures for Emotion Recognition," Neural Networks, vol. 92, 2017.

[12] M. N. Sttilar, M. Leich, R. S. Bolie, M. Skinter, "Real-time emotion recognition using RGB image classification and transfer learning," 11th Int. Conf. Signal Process. Commun. Syst., 2017.

[13] Q. Maio, M. Diong, Z. Huaeng, Y. Zhaen, "Learning Salient Features for Emotion Recognition Using Neural Networks," IEEE Trans. Multimedia., vol. 16, 2014.

[14] P. Hairár, R. Burgeet, M. K. Duitta, "Facial Emotion Recognition with Deep Learning," in Signal Processing and Integrated Networks (SPIN), 2017.

[15] Nithya Roopa. S. "Emotion Recognition from Facial Expression using Deep Learning," International Journal of Engineering and Advanced Technology (IJEAT), August 2019.

[16] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, S. Narayanan, "Analysis of Emotion Recognition using Facial Expressions, Speech and Multi-modal Information", ICMI'04, October, 2004.

[17] Massaro, D. W. "Illusions and Issues in Bimodal Speech Perception", Proceedings of Auditory Visual Speech Perception '98., December, 1998.

[18] De Silva, L. C., Miyasato, T., and Nakatsu, R. "Facial Emotion Recognition Using Multi-modal Information", In Proc. IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS'97), Sept. 1997.

[19] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G. "Emotion recognition in human-computer interaction". Signal Processing Magazine, IEEE, Volume: 18, Issue: 1, Jan 2001.

[20] Pantic, M., Rothkrantz, L.J.M. "Toward an affect-sensitive Multi-modal human-computer interaction". Proceedings of the IEEE , Volume: 91 Issue: 9 , Sept. 2003.

[21] Roy, D., Pentland, A. "Automatic spoken affect classification and analysis. Automatic Face and Gesture Recognition", Proceedings of the Second International Conference, Oct. 1996.

[22] Dilbag Singh. "Human Emotion Recognition System", International Journal of Image Graphics and Signal Processing, August 2012.

[23] R. D. Berger, S. Akselrod, D. Gordon, R. J. Cohen, "An efficient algorithm for spectral analysis of heart rate variability", IEEE Trans. Biomed. Eng., 33, pp. 900–904.

BIOGRAPHIES

RUTUJA TAMBE
Student at AISSMS COE, Pune-01.
Savitribai Phule Pune University,
B.E. in Computer Science.

ANMOL THETE
Student at AISSMS COE, Pune-01.
Savitribai Phule Pune University,
B.E. in Computer Science.

ANIKET GADHAVE
Student at AISSMS COE, Pune-01.
Savitribai Phule Pune University,
B.E. in Computer Science.

SUMEDH DHENGRE
Assistant Professor at AISSMS COE, Pune-01.
Savitribai Phule Pune University,
M.E. in Computer Science