Emotion Analysis of Students

Amanda Judy Andrade Komal Kadam Nilam Kadam Vaishali Kavathekar
Department of Information Technology, Don Bosco Institute of Technology, Affliated to University of Mumbai, Premeier Automobile Road, Kurla-West, Mumbai-400070
Corresponding Author: Vaishali Kavathekar (vaishalik.dbit@dbclmumbai.org)

***Abstract*— Mental health organizations embedded inside instructive frameworks can make a continuum of integrative thought that improves both passionate prosperity and informative accomplishment for youngsters. To strengthen this continuum, and for ideal child improvement, a reconfiguration of preparing and passionate health structures to help execution of verification-based practice might be required. Integrative methods that merge study lobby level and understudy level interventions have a ton of possibilities. Summarizing the proposed procedure by testing it in group and social learning stages, an unpretentious understudy responsibility examination can be used to make shrewd instructing structures more modified.**

Keywords*— *Emotion Recognition, Convolution Neural Network, Facial Action Coding System, Facial Action Unit.

I. INTRODUCTION

The emotional health of the student and the overall performance in Academics, Extra-Curricular Activities are directly correlated to each other. Government bodies allocate funds for education and [1] [2] human welfare in most of the developed and developing countries, to enhance the economy of the country, this indicates they require a prolific educated workforce to act on demand. Tending to the issue of the student's emotional health by distinguishing it at the beginning phase and following up on the necessary changes can assist with expanding the overall performance.

It tends to be hard for instructors to distinguish tension and despondency because these issues frequently show up diversely for various individuals, however, this is the reason knowing the blends of practices to search for is critical. An understudy managing one of these issues can encounter negative consequences for their consideration, translation, focus, memory, social interaction and actual wellbeing. At the point when somebody is encountering uneasiness or wretchedness most of their intellectual ability is utilized to make and deal with troubling contemplations. This can make it amazingly hard to zero in on certain musings and can be exceptionally debilitating for the understudy, which takes away from their learning capacities.

Psychological wellness administrations inserted inside educational systems can make a continuum of integrative consideration that improves both emotional well-being and instructive achievement for kids. To fortify this continuum, and for ideal kid improvement, a reconfiguration of training and emotional wellness frameworks to help execution of proof-based practice may be required. Integrative techniques that consolidate study hall level and understudy level intercessions have a lot of potentials. A strong exploration plan is required that centers around framework level execution and support of intercessions over the long run. Both moral and logical avocations exist for a mix of emotional wellness and training: coordination democratizes admittance to administrations and, whenever combined with utilization of proof-based practices, can advance the sound improvement of kids.

Summing up the proposed strategy by testing it in collective and social learning stages, a subtle understudy commitment investigation can be utilized to make wise coaching frameworks more customized.

II. RELATED WORK

The authors, A. Sharma and V. Mansotra; [3] conducted their research based on facial emotion recognition of students in a classroom arrangement and have proposed a deep learning approach to analyse emotions with improved emotion classification results and offer optimized feedback to the instructor. A deep learning-based convolution neural network algorithm was used in this paper to train FER2013 facial emotion images database and they used transfer learning technique to pre-train the VGG16 architecture-based model with Cohn-Kanade (CK+) facial image database, with its own weights and basis. A trained model captured the live steaming of students by using a high-resolution digital video camera that faces towards the students, capturing their live emotions through facial expressions, and classifying the emotions like sad, happy, neutral, angry, disgust, surprise, and fear, it offered the authors insight into the class group emotion that is reflective of the mood among the students in the classroom. This experimental approach can be used for video conferences, online classes etc. They presented their research methodologies and their achieved results on student emotions in a classroom atmosphere and had proposed an improved CNN model based on transfer learning that could suggestively improve the emotions classification accuracy. Krithika L.B & Lakshmi Priya GG, [4] proposed a system that would detect emotions based on Gaussian distance between the eyes and eyebrows. It eliminated the need of any device usage requiring physical contact to the subject under study. The existing system helped to identify emotions and classify learner involvement and interest in the topic were plotted as feedback to the instructor to improve the learning experience. This serves as a stepping stone to our proposed system. Experiments, (J. Guo et al; [5]) indicated that pairs of compound emotion (e.g., surprisingly-happy vs happily-surprised) were more difficult to be recognized if compared with the seven basic emotions. The recognition of compound emotions on the iCV-MEFED dataset demonstrated to be very challenging, leaving a large room for improvement. Top winners' methods from FG 2017 workshop have been analysed and compared. How to incorporate prior information of dominant and complementary categories into compound facial emotion recognition is one question we want to address in future work. The authors (T. S. Ashwin and R. M. R. Guddeti) proposed a convolutional neural network[6] architecture for unobtrusive students' engagement analysis

using non-verbal cues. The proposed architecture was trained and tested on faces, hand gestures and body postures in the wild of more than 350 students present in a classroom environment, with each test image containing multiple students in a single image frame. The data annotation was performed using the gold standard study, and the annotators reliably agree with Cohen's κ = 0.43. They obtained 71% accuracy for the students' engagement level classification. Further, a pre-test/post-test analysis was performed, and it was observed that there is a positive correlation between the students' engagement and their test performance. This existing research is considered for our proposed system. The authors (Fennell, P.G., Zuo, Z. & Lerman, K) [7] described a statistical approach to modelling behavioural data called the structured sum-of-squares decomposition (S3D). The algorithm, which was inspired by decision trees, selects important features that collectively explain the variation of the outcome, quantifies correlations between the features, and bins the subspace of important features into smaller, more homogeneous blocks that correspond to similarly-behaving subgroups within the population. They proved that S3D creates parsimonious models that can predict outcomes in the held-out data at levels comparable to state-of-the-art approaches, but in addition, produces interpretable models that provide insights into behaviours. This is important for informing strategies aimed at changing behaviour, designing social systems, but also for explaining predictions, a critical step towards minimizing algorithmic bias. The authors Y. Tang, Q. Mao, H. Jia, H. Song and Y. Zhan; proposed an emotion-embedded visual attention model (EVAM) [8] to learn emotion context information for predicting affective dimension values from video sequences. First, deep CNN was used to generate a high-level representation of the raw face images. Second, a visual attention model based on the gated recurrent unit (GRU) was employed to learn the context information of the feature sequences from facial features. Third, the k-means algorithm was adapted to embed previous emotion into attention model to produce more robust time-series predictions, which emphasize the influence of previous emotion on current effective prediction. In this paper, all experiments were carried out on database AVEC 2016 and AVEC 2017. The experimental results validate the efficiency of the proposed method, and competitive results were obtained. We consider this project as an important guiding stone for our proposed system. In this paper (Q. Mao, Q. Zhu, Q. Rao, H. Jia and S. Luo), a novel three-stage method [9] was proposed to learn hierarchical emotion context information (feature-and label-level contexts) for predicting affective dimension values from video sequences. In the first stage, a feed-forward neural network was used to generate a high-level representation of the raw input features. Then, in the second stage, the bidirectional long short-term memory (BLSTM) layers learn the context information of the feature sequences from the high-level representation and get the initial recognition results of the input. Finally, in the third stage, a BLSTM neural network was used to learn the context information from emotion label sequences by an unsupervised way, which was used to correct the initial recognition results and get the final results. The authors explored the influence of different sequence lengths by sampling from the original sequences. The experiment performed on the video data of AVEC 2015 demonstrated the effectiveness of the proposed method. Their framework highlights that incorporating both feature/label level dependencies and context information is a promising research direction for predicting the continuous dimensional emotion. This research is a stepping stone to our proposed system as it tells us which dataset can help the us in getting a better yield.

III. Proposed Methodology

The exploration started by investigating existing examination papers and characterizing our difficult assertion. Whenever this was accomplished could we have gone to a remain of choosing a last strategy for execution. The creators were proposed to have a go at carrying out the exploration on a solitary individual to identify whether the calculation works in the correct way. In any case, on conversation with the board individuals they needed to utilize FER2013 dataset. They have effectively executed the proposed procedure in the ReLU 4-layered 2D R-CNN Design with the calculation's learning rate (**α**) to 0.0005, β_1=0.9, β_2=0.999, and epsilon ε=e-5 with the precision of the model at 66.7% and lower misfortune pace of 89% more than 50 epochs, in this way astounding past models.

A. Dataset Collection and Preprocessing

The dataset comprises 48x48 pixel grayscale pictures of countenances [10]. The appearances have been consequently enrolled with the goal that the face is pretty much focused and possesses about a similar measure of room in each picture. The errand is to order each face dependent on the feeling that appeared in the outward appearance into one of seven classes (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

Figure 1:Samples from FER2013 Dataset

The substance of this string a space-isolated pixel esteems in line with significant request. The preparation set comprises 28,709 models. The public test set utilized for the leader board comprises 3,589 models. The last test set, which was utilized to decide the victor of the opposition, comprises of other 3,589 models.

Preprocessing of data includes grayscale conversion of RGB images, face detection and crop, image normalization and image augmentation.

B. Convolution Neural Network (CNN):

Convolution Neural Network is a popular method for image classification that has been chosen with data augmentation in this research.

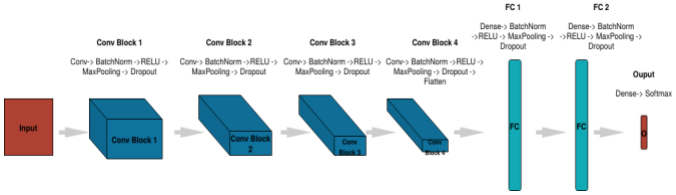

Figure 2: Architecture of the Proposed Model

First of all, images from the dataset were fed into the face detection system using the Haar Cascade Classifier to detect face in the input image. If faces are detected from the visual input, the input is passed to ImageDataGenerator function from Keras API for further data augmentation such as flipping, rotating, shearing etc. The face then passes into the CNN classifier to predict classes.

The proposed model used to classify facial expression consists of 4 convolution layers with 64, 128, 512 and 512 filters respectively with the 3x3 kernel. Each layer consists conv2D layer, max-pooling layer, dropout layer, and batch-normalization layer. The conv2D is used to specify convolution kernel, which is 3x3 kernel. The max-pooling layer is used for dimension reduction by finding the maximum value in the 2x2 windows. The dropout layer is included to avoid overfitting problem. The training time optimization is performed by the batch-normalization layer. The activation function using in all layer except the output layer is Rectified Linear Unit (ReLU) activation function. The ReLU is a function aiming to faster converge cost to zero and boost accurate results. The Categorical Cross-entropy loss is used as a loss function in the output layer. This categorical cross-entropy loss is a combination of the Softmax activation and Cross-entropy loss functions used for multi-class classification to distinguish output into 7 classes of emotion.

C. Face Detection Algorithm-Viola Jones:

The job of each stage [11] is to determine whether a given sub-window is definitely not a face or may be a face. A given sub-window is immediately discarded as not a face if it fails in any of the stages.

A simple framework for cascade training is given below:

- f = the maximum acceptable false positive rate per layer.
- d = the minimum acceptable detection rate per layer.
- F_{target} = target overall false positive rate.
- P = set of positive examples.
- N = set of negative examples.

The cascade architecture has interesting implications for the performance of the individual classifiers. Because the activation of each classifier depends entirely on the behavior of its predecessor, the false positive rate for an entire cascade is: $F = \prod_{i=1}^{K} f_i$. Similarly, the detection rate is: $D = \prod_{i=1}^{K} d_i$.

D. Adam Optimizer

Adam [12] is a versatile learning rate technique, which implies, it figures singular learning rates for various boundaries. Its name is gotten from adaptive moment estimation, and the explanation it's called that will be that Adam utilizes assessments of the first and second moments of the gradient to adjust the learning rate for each weight of the neural organization.

N^{th} moment of a random variable is defined as the expected value of that variable to the power of n, i.e.

$$m_n = E[X^n] \qquad (1)$$

To estimates the moments, Adam utilizes exponentially moving averages, computed on the gradient evaluated on a current mini-batch:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \quad (2)$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \quad (3)$$

Where m and v are moving averages, g is gradient on current mini-batch, and betas — new introduced hyper-parameters of the algorithm. They have really good default values of 0.9 and 0.999 respectively. Almost no one ever changes these values. The vectors of moving averages are initialized with zeros at the first iteration.

Weight Update in Adam occurs in following manner with η denoting the step-size:

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \qquad (4)$$

where,

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \qquad (5)$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \qquad (6)$$

Equations 5 and 6 indicate a bias correction for the estimators. Results obtained from the above equations are used in weight update.

IV. Experiment and Result

The experiment was conducted over a sample of stored and real-time videos. The model was deployed and obtained the following results:

A. Image Augmentation

A good classifier requires loads of training dataset to accomplish a good outcome. Image Augmentation is an interaction to assist with artificial training images using various methods of image processing. Flipping, pivoting, cropping, shading jittering, edge upgrade and extravagant PCA are famous augmentation techniques. In this paper, cropping, pivot, shear, zoom and flip have been utilized to improve the precision of the model.

B. Model Implementation

The emotion classification model has been written in the python programming language with Keras, TensorFlow, NumPy, PIL, OpenCV, and Matplotlib libraries. The Keras provides activation function, optimizers, layers, dropout, batch normalization, etc. The TensorFlow was used as a system backend to accept the inputs of a multidimensional array, which are the pixels of trained images. The OpenCV was used mainly to detect a face in the image or video streaming using Haar Cascade classifier, grayscale image conversion, and image normalization. The graphic user interface (GUI) was written in a python programming language to accept both still image and real-time video streaming. The system, then, converts the input into 48*48 grayscale image after the face is detected by Haar Cascade Classifier. After that, the cropped image has been passed into the proposed model to classify to 7 distinct emotions.

C. Overall Recognition Accuracy

In the experiment, the accuracy and loss assessment of the proposed model against the state-of-the-art model, which is VGG-16 in the task of emotion classification based on FER2013 dataset has been performed. The optimized Convolutional Neural Network (CNN) with additional implementation of image augmentations, layer dropout and normalization with less complexity of 5,786,247 parameters has proven to be more efficient compared to the state-of-the-art model with higher complexity of 5,790,727 parameters with the same dataset and numbers of training epochs by

achieving a higher rate of model accuracy at 66.70% validation accuracy at 64.50% with the shorter training time at 50 epochs as shown below.

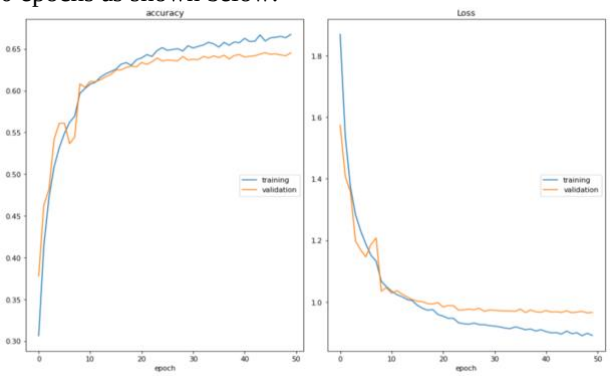

Figure 3: Model Accuracy and Model Loss of Proposed Model used

Table 1:Model Loss and Accuracy Data

Data/Metrics	Accuracy	Loss
Training	66.70%	89.1%
Validation	64.50%	96.6%

V. Results

The following figures depicts the results of emotion detected real-time and stored videos (CCTV, news roll TV). Additionally, utilizing combined models is likewise a fascinating examination to improve model precision and diminish training time.

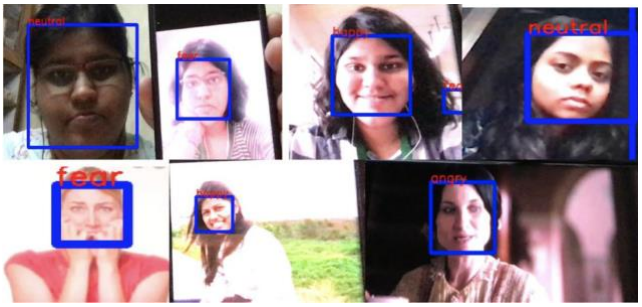

Figure 4:Real-Time Emotion Recognition

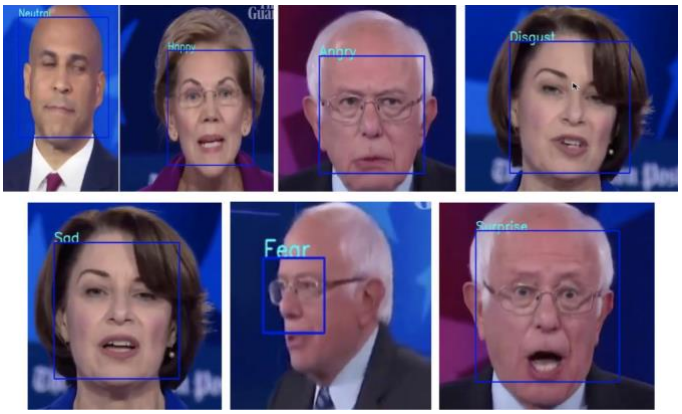

Figure 5:Emotion Detection over Stored Video. Presidential Debate 2020(https://youtu.be/BIm_KpBshtA)

VI. Discussion

Moreover, the trained model can be deployed in a Mobile Application using TensorFlow Lite and Firebase for identifying emotions on mobile devices through stored videos or through their device cameras. Amazon Rekognition can be used as an alternative for upcoming project for face recognition.

References

[1] International Board of Credentialing and Continuing Education Standards, "Impact of Anxiety and Depression on Student Academic Progress," International Board of Credentialing and Continuing Education Standards, [Online]. Available: Available: https://ibcces.org/blog/2019/05/01/impact-anxiety-depression-student-progress/.

[2] H. K. S. S. F. T. F. M, "Mental Health Interventions in Schools in High-Income Countries," *Lancet Psychiatry*, vol. 1, no. 5, pp. 377-387, 2014.

[3] V. M. A. Sharma, "Deep Learning based Student Emotion Recognition from Facial Expressions in Classrooms," *nternational Journal of Engineering and Advanced Technology (IJEAT)*, vol. 8, no. 6, p. 2249 – 8958, 2019.

[4] L. P. G. K. L.B, "Student Emotion Recognition System (SERS) for e-learning Improvement Based on Learner Concentration Metric," in *International Conference on Computational Modeling and Security (CMS 2016)*, 2016.

[5] J. G. e. al., "Dominant and Complementary Emotion Recognition From Still Images of Faces," *IEEE Access*, vol. 6, no. DOI: 10.1109/ACCESS.2018.2831927, pp. 26391-26403, 2018.

[6] T. S. A. a. R. M. R. Guddeti, "Unobtrusive Behavioral Analysis of Students in Classroom Environment Using Non-Verbal Cues," *IEEE Access*, vol. 7, no. DOI: 10.1109/ACCESS.2019.2947519, pp. 150693-150709, 2019.

[7] P. Z. Z. &. L. K. Fennell, "Predicting and explaining behavioral data with structured feature space decomposition," *EPJ Data Sci*, vol. 8, no. https://doi.org/10.1140/epjds/s13688-019-0201-0, 2019.

[8] Q. M. H. J. H. S. a. Y. Z. Y. Tang, "An Emotion-Embedded Visual Attention Model for Dimensional Emotion Context Learning," *IEEE Access*, vol. 7, no. DOI: 10.1109/ACCESS.2019.2911714., pp. 72457-72468, 2019.

[9] Q. Z. Q. R. H. J. a. S. L. Q. Mao, "Learning Hierarchical Emotion Context for Continuous Dimensional Emotion Recognition From Video Sequences," *IEEE Access,*, vol. 7, no. DOI: 10.1109/ACCESS.2019.2916211, pp. 62894-62903, 2019.

[10] [Online]. Available: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data..

[11] J. Viola, "Robust Real-time Object Detection," in *International Journal of Computer Vision*, 2001.

[12] D. P. K. a. J. L. Ba, "Adam : A method for stochastic optimization," no. arXiv:1412.6980v9, 2014.