ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, *15-17 June*, *2019*, *pp. 265-273*
doi: 10.18178/wcse.2019.06.040

Deep Learning Approach for Identifying Emotions in IELTS Speaking Tests

Lenin Kahanga[+], Yan Wang[+]

College of Computer Science and Technology, Harbin Engineering University, Harbin, China
leninkaha@gmail.com, wangyanj@hrbeu.edu.cn

Abstract. This paper proposes a novel deep learning framework to identify student emotions or affective states in IELTS speaking video tests. This approach has one unique characteristic, it extracts features from a face image and sends it to a deep neural network model for emotion classification. The main objective of this study is to find the correlation between the test takers emotion status and their test grades. This framework is evaluated with extensive experiments. The achieved results show promising performance based on the size of the data and computing resources. The outcomes of this work would add value to OEP systems.

Keywords: Deep learning, Proctoring, OEP

1. Introduction

Human communication conveys content and attitude. Emotional aspects play an important role in understanding communication, decision making and behavioral aspects. Voice is a verbal form of communication while facial expressions, body posture and gestures constitute non-verbal communication. While communicating only 7% effect of message is contributed by verbal part as a whole, 38% by vocal part and 55% effect of the speaker‟s message is contributed by facial expression as in [4]. Therefore real time facial expression analysis plays an important role in determining human emotions/feelings.

According to [4], the extraction of emotion from the static image allows the recognition of several physical features such as: eyes, wrinkles on the forehead, size of eyebrows, color of the skin and the corresponding sizing and location. In this case, the neural network is accurate for the acquisition of nonlinear mapping between different sets of data, this analysis allows decode the relationship between the physical features of a face and its impression. The potential of the neural-network-based methods is the performance of facial expression classification into a single basic emotion category.An application of artificial neural network (ANN), deep learning and support vector (SVM) methodologies is used [5], to identify personality traits using facial images. The personality traits studied are being, attractive, caring, aggressive, intelligent, stable, confident, trustworthy, dominant, unhappy. In [6], a deep learning based approach has been demonstrated for the task of semantic facial feature recognition. This approach is primarily based on the use of convolutional neural networks on two dimensional pre-processed and aligned images of faces. In the experiments emotion classification, age classification, gender classification, ethnicity classification was performed at different stages. The study suggests that a deep convolutional network based approach is naturally well suited for the task of image based facial expression recognition[9]. Such a deep network can easily be adapted to the tasks of recognizing additional semantic features[10]. Experimental results showed near-human performance

[+] Corresponding author. Tel.: 13144626853
E-mail address: leninkaha@gmail.com; wangyanj@hrbeu.edu.cn

Massive open online courses (MOOCs) or e-learning offer the potential to significantly expand the reach of today‟s educational institutions, both by providing a wider range of educational resources to enrolled students and by making educational resources available to people who cannot access a campus due to location or schedule constraints. Instead of taking courses in a typical classroom on campus, now students can take courses anywhere in the world using a computer, where educators deliver knowledge via various types of multimedia content[2]. According to a recent survey [1][2], more than 7.1 million students are taking, at least, one online course in 2013 in America. It also states that 70% of higher education institutions believe that online education is a critical component of their long-term strategy [2]. The rapid growth of e-learning has created various supporting technologies. One of these is online or virtual proctoring, the monitoring of a candidate during an exam through use of webcam, mic or the computer screen. Yousef Atoum et al [3], proposes an analytics system to perform automatic and continuous online exam proctoring (OEP). The overall goal of this system is to maintain academic integrity of exams, by providing real-time proctoring for detecting the majority of cheating behaviors of the test. To achieve such goals, audio-visual observations about the test takers are required to be able to detect any cheat behavior [3]. Their work is motivated by previous research which has utilized audio-visual features to study human behavior. The system monitors such cues in the room where the test taker resides, using two cameras, a microphone and proposes a hybrid two-stage algorithm for our OEP system. The first stage focuses on extracting middle-level features from audio-visual streams that are indicative of cheating [3]. These mainly consist of six basic components: user verification, text detection, speech detection, active window detection, gaze estimation, and phone detection. According to [13] its discussed that student affective states such as interested, tired, and confused can be determined from facial expressions and attention state is computed from different visual cues such as face gaze, head motion, and body postures.Nigel Bosch et al [7], demonstrated that automatic detection of boredom, confusion, delight, engagement, and frustration in natural environments was possible for students using an educational game in class despite the many real-world challenges for these classification tasks, such as classroom distractions and large imbalances in affective distributions [7]. With respect to class distractions, students in the current study fidgeted, talked with one another, asked questions, left to go to the bathroom, and even occasionally used their cellphones.

While prediction of emotions has been reported in the above literature, most of the work focusses on cheating and attention. There is a challenge of identifying emotions during tests. To the best of our knowledge there is no prior work in identifying student emotions during tests. In this paper a novel deep learning and emotion detection method for IELTS speaking tests is presented. In addition to emotion classification the association between the test takers grades and their emotion status during the test is studied. This work will help improve the learning process, for instance, positive emotions mostly increase student‟s interest, engagement and motivation to meet their goals, while anxious, angry and depressed students don‟t perform well. This paper is organized as follows, chapter II gives the general methodology and tools to be used, chapter III covers the experimental evaluation and discussion, in chapter IV we draw conclusions and point towards future work.

2. Proposed Methodology

The proposed framework is presented in this chapter. It has three sections: first section gives a description of the required features, the second describes the dataset to be used and in the last section the model design is given.

2.1. Feature selection

Emotions are part of every human interactions and thus their identification and analysis can help us improve the understanding of different human actions. Although a variety of human emotions exist. According to Ekman [1992] psychologists focus on 8 primary emotions namely. Anger, Fear, Happiness, Sadness, Disgust, Surprise, Trust and Anticipation. Monika Dubey et al [4] discusses the primary emotions and their breakdown into secondary emotions. Furthermore, they describe the facial movements for the different emotions. From this literature we discover human affective states are quite many and complex thus not easy to study fully. Nigel Bosch et al [7] identifies the five most common affective or emotional states;

boredom, confusion, engagement/flow, happiness/delight and anxiety which will be mainly discussed in this paper.From the various myriad blends of emotions discussed, previous literature by Nabeela Altrabsheh et al [8], identifies the main emotions relevant to learning; Bored, Amused, Frustration, Excitement, Enthusiasm, Anxiety, Confusion, and Engagement. We use these emotions to define an emotion classification that will be the focus of this research and will be used in this task. The classification is shown in table "Table 1" below.

Table 1

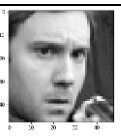
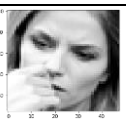
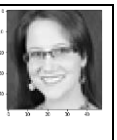
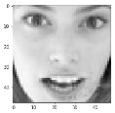
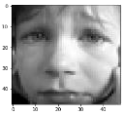
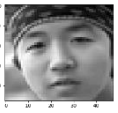
Emotion	**Tertiary Emotions**
Amused	Cheerfulness, elation, content, delight, joy
Frustrated	Anger, dislike
Bored	Disgust, loath
Excited	Enthusiasm, thrill, zeal
Anxiety	Fear, nervousness, panic, fright
Confused	Sadness, disappointment
Engaged	Trust, acceptance

2.2.Dataset description

The 2 datasets to evaluate the proposed approach are sourced from youtube and kaggle. Both were found to be well suited to extract the dataset because most of it is available free and contains diversity.

The first dataset (Image data) was collected from the Kaggle competition dataset prepared by Pierre-Luc Carrier and Aaron Courville. The dataset consists of 38888 images divided as training (28709) and testing (3589). The images are 48x48 pixel grayscale images of individual faces from a cross spectrum of race, gender and ethnicity. The faces are centered and occupy the same amount of space in each image. The features are stored in csv file with 2 columns „pixels" and „emotions. The pixel column contains the pixel values of the image. The emotions column contains a numeric value ranging from 0-6 representing the emotion present in the image. These are emotion categories represented (0 for Angry, 1 for Disgust, 2 for Fear, 3 for Happy, 4 for Sad, 5 for Surprise, 6 for Neutral). Below are sample pictures generated from the dataset "Table 2"

Table 2

			
angry/frustrated	disgust/bored	fear/anxiety	happy/amused
			
surprise/excited	sad/confused	neutral/engaged	

The second dataset (video dataset) was extracted from social media site youtube. Beautiful Soup a python-based web scrapping library was used. Search queries focused on sample IELTS speaking test videos because they relate to the learning environment. To maintain data integrity, only videos from 2 recognized youtube channels. The official IELTS youtube channel (IELTS Official) and the Reading IELTS youtube channel. The final video dataset consists 13 videos sourced from the two channels. All the students are of different gender, race, sex and they express themselves in English language. The video length ranges from 4-5 minutes.

For this research, all the videos were converted to mp4 format of dimension 360 x460. The length of the videos varies from 2 to 3 minutes. Most youtube videos contain introductory text such as title or company logo. The videos are preprocessed to remove the introductory text and, in this work, the first 5 seconds of each video were removed. There are two people in each video the test administrator and the test taker. The test administrator asks the questions and the test taker gives the answers. All the videos as segmented to remove the parts of the test administrator, only the parts of the test taker are maintained

2.3. Model Design

The CNN model processing pipeline is a series of: convolution, max-pooling and a fully connected network layers. In general, a convolution is a result of two functions of a real value "Fig. 1." It gives a weighted average of all measurements of a given real value, often denoted as "(1)",

$$s(t) = (x * w)(t) \tag{1}$$

where s is output, x input and w is the kernel

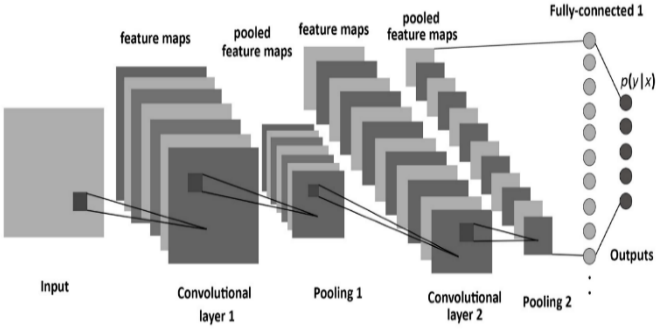


Fig. 1: CNN- Convolutional Neural Network

The first argument of a convolution, in this case function x is called the input, the second argument function w is referred to as the kernel, the output of the function is referred to as the feature map. The input to a convolutional network is always a multidimensional array of data and the kernel is also a multidimensional array of parameters adapted by the algorithm, these multidimensional arrays are referred to as tensors. Because each element of the input and kernel must be explicitly stored separately, we usually assume that these functions are zero everywhere but in the finite set of points for which the values [3][11] are stored. This means that in practice, implement the infinite summation as a summation over a finite number of array elements. Finally, use convolutions over more than one axis at a time [3]. If a two-dimensional image is used as input, probably also use a two-dimensional kernel K (2):

$$S(i,j) = (I * K)(i,j) = \sum_m m \sum_n n\, I(m,n) K(i - m, j - n). \tag{2}$$

First the layer performs several convolutions to produce linear activation functions, then the linear activation functions are run through non-linear functions such as ReLU and finally the apply a pooling function to modify the output of the layer. A pooling function replaces the output of the net at a certain location with a summary statistic of the nearby outputs. For example, the max pooling [3][12] operation reports the maximum output within a rectangular neighborhood. Pooling is essential for handling inputs of varying size. For instance, if the goal is to classify images of variable size, the input to the classification layer must have a fixed size. This is usually accomplished by varying the size of an offset between pooling regions so that the classification layer always receives the same number of summary statistics regardless of the input size. For example, the final pooling layer of the network may be defined to output four sets of summary statistics, one for each quadrant of an image, regardless of the image size [3].

The key function of the CNN that determines which values to pick, how to assign weights during training is a process called backpropagation. Backpropagation is made up of 4 stages namely; forward pass, loss function, backward pass and weight update. Forward pass, you take a training image in the form of an array of numbers and send it to a network. To determine which weights contributed to the loss and change them so as to reduce the loss, perform a backward pass. After computing the backward pass, perform the weight update. Here the weights of all the filters are updated so that they change in the opposite direction of the gradient as show in formula (3);

$$W = w0 - r\frac{dL}{dW} \tag{3}$$

where W is weight, w0 as the initial weight, r learning rate and L loss.

The learning rate is a random number set during model design, high learning rate implies big changes in weight updates and the model is likely to take less time to converge on an optimal set of weights. The combination of forward pass, loss function, backward pass, and weight update is one training iteration. This

process will be repeated on a fixed number of images from the training set (known as a batch). By the end of this process the network is well trained that the weights will be tuned correctly.

To ensure that the algorithm works well on not just on training data but also on new inputs, apply regularization. It works by estimator regularization. Estimator regularization is done through trading reduced variance for increased bias. The most effective regularizer makes a tradeoff between reducing variance, while not overly increasing the bias. Use dropout regularization which removes some nodes from the network [3]. The dropout uses a minibatch-based learning algorithm that makes small steps, such as stochastic gradient descent. Every time a sample is added into a minibatch, randomly sample a different binary mask to apply to all the input and hidden units in the network [3]. The mask for every unit is sampled independently from the others. The probability of sampling a mask value of one (causing a unit to be included) is a hyper parameter fixed before training begins [3]. Suppose a vector μ which defines which units to include and $J(\theta, \mu)$ the model cost defined by parameters θ and μ.Dropout consists minimizing$E_{\mu}J(\theta, \mu)$. In this case for the dropout, each sub model defined by mask vector μ defines a probability distribution $p(y|x, \mu)$, the mean over all masks is given by"(4)"

$$\sum_{\mu} p(\mu)p(y|x, \mu) \tag{4}$$

where p (μ) is the probability distribution that was used to sample μ during training.

3. Experiments and Results

In this section we use a series of experiments to investigate the proposed deep learning model. The model is designed using Keras, a deep learning python library that runs on top of Tensorflow, an open source numerical computing library developed by Google. Keras was used because of its speed and easier implementation with minimal computing resources.

3.1. CNN model

This model was trained and tested with images from the FER-2013 dataset. The dataset was collected from kaggle. The dataset represents the 7 common emotions on every item in the dataset. The pixel information extracted from each facial image is stored in a csv file. The visual features are extracted and classified using a deep convolutional neural network (CNN). The network receives an image input and analyses the image features.CNN applies a supervised learning approach on huge image datasets. The CNN pipeline used identify the emotion in a facial image is shown in "Fig. 2."

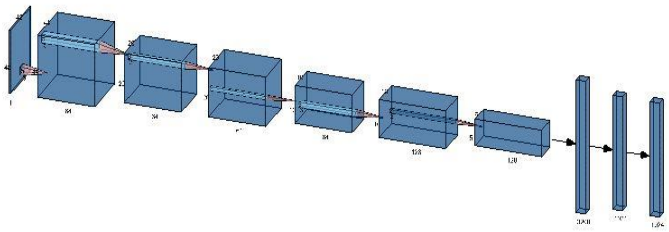


Fig. 2: Traning model(CNN)

The network is composed of 3 convolution layers, 2 dense layers, 1 output layer, ReLU activation functions and pooling.

The input layer is fed with pixel values of an image as (x, y, z) where x represents image width, y represents image height and z is the number of colors. For this model (48, 48, 1) is used where 1 represents grayscale images.

Before the pixels are fed into the input layer, the values are preprocessed to fit into the fixed dimensions.

The convolutional layer calculates the dot product of the weights and a tiny image tile to which the neurons are connected in the input layer. The number of tiles is input as hyper parameters with unique randomly generated weights. For the max pooling size (2, 2), kernel sizes (3, 3) and (5, 5) are used. For

accuracy a dropout rate of 0.3 and batch normalization are used. We apply ADAM optimization, 250 batches and 5epochs to avoid overfitting.

3.2. Training results

The model yields an accuracy of 60%. This is promising based on the computing resources used. The accuracy can be improved with more training on GPU resources. Below plot “Fig. 3.” shows the history of loss and accuracy metrics at the successive epochs.

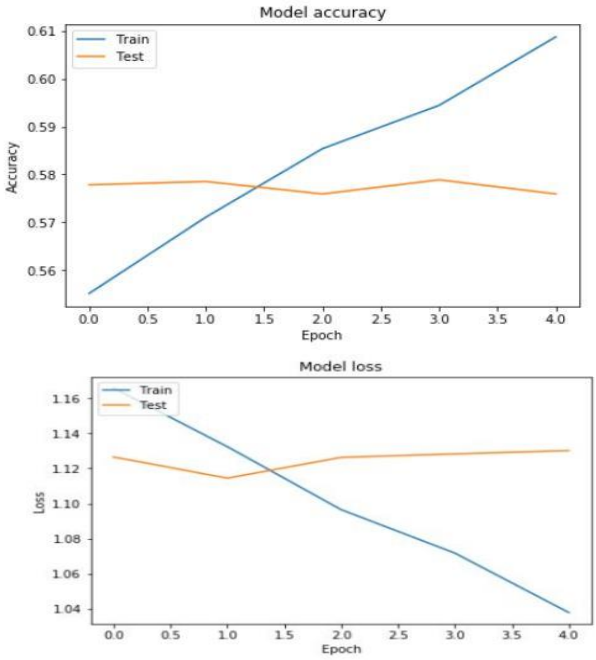


Fig. 3: Training and testing accuracy/loss plot.

From the experiment as you can see in the plot above, the training accuracy increases with every epoch and the training loss goes down at every epoch.

In addition, a confusion matrix metric is used to evaluate the model as given in figure “Fig. 4.” Rows represent the actual values and the columns represent the model predictions. Example, there are 376 frustration instances (column 1) in the test dataset and 184 are classified correctly. 27 are classified as bored, 43 classified as anxious, 19 amused, 56 confused, 8 excited and 39 engaged, yet they are all frustration instances.

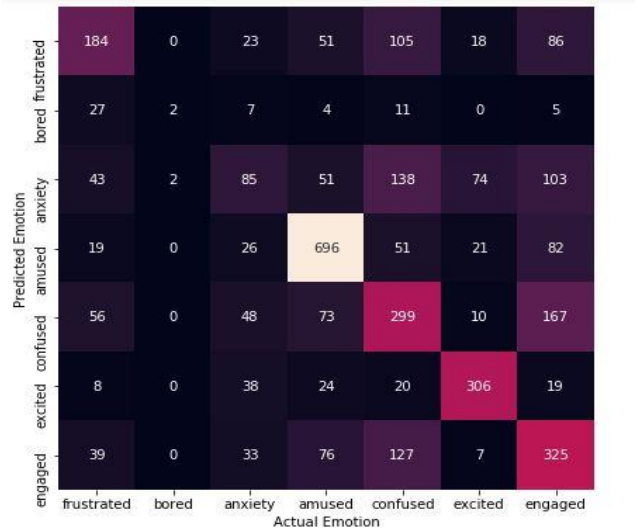


Fig. 4: Confusion matrix

3.3. Model validation

In this section, validation of the proposed framework by applying it on the IELTS video dataset is done.

Under this approach the input video is processed to extract video frame sequences. OpenCV haar feature-based cascade classifiers is used to detect the human face in each frame. The haarcascade takes facial features from every frame and averages them. The features are used as input to the CNN model, which is the core of this approach used to classify the emotions in a frame. A detailed illustration of this approach is shown in the figure "Fig. 5."

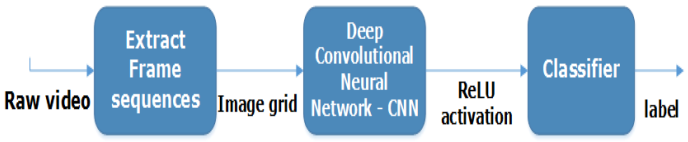


Fig. 5: System design to identify emotions in video

The steps below are followed to implement this approach.

Each video is segmented into several frames according to the frame rate of the video

Then extraction of the facial features from each video frame and compute the average to make a final feature vector which is a representation of the whole video

Use the final feature vector to classify the flow of emotions in the video in percentage terms.

The results are given in table "TABLE 3." below

Table 3

video	**Emotion (%)**						
	frustrated	bored	anxiety	amused	confused	excited	engaged
Vid1	5	2	8	**40**	15	3	25
Vid2	6	0	11	21	**33**	3	23
Vid3	11	1	8	7	26	2	**42**
Vid4	18	1	17	5	**42**	2	18
Vid5	**34**	1	21	4	13	8	8
Vid6	10	1	10	25	**33**	2	22
Vid7	24	0	4	11	20	0	**40**
Vid8	10	0	6	**33**	23	3	23
Vid9	8	0	14	7	**43**	2	25
Vid10	8	1	25	2	**33**	2	10
Vid11	8	0	3	**63**	9	1	11
Vid12	14	1	15	7	**36**	2	12
Vid13	10	0	4	**74**	5	2	5

From the results the predominant emotion with the highest percentage (blue color) is taken as the emotion classification of the video. Emotion detection is a pivotal component for learning environments, that aspire to improve and understand students learning process and engagement by responding to the observed emotion. This paper advances research in this field by demonstrating an emotion identification model for IELTS speaking tests. In previous research related to identifying emotions in the academic environment focused on textual feedback. This work demonstrated that detecting these emotions; amused, confused, engaged,

frustrated and anxiety for students during speaking tests is possible. We observe that videos of students with high test grades (8 and 9), portrayed high percentages for positive emotions like engaged and amused. The videos with high percentages of negative emotions such as frustrated, anxious, confused crossed average/lower test grades between 5,5.5and 6. From this observation, a conclusion is drawn that there is strong association between test scores and emotion status during the test. Whereas it is difficult to compare this model‟s performance with past works due to differences in data and emotions considered the technique proposed in this work provides a robust method to recognize emotions during video tests and can be scalable to other samples such as interviews.

4. Conclusion

In conclusion, we proposed a deep learning model to detect emotions during IELTS speaking tests trained and tested on standard datasets sourced from kaggle and youtube. Deep convolutional neural networks are employed to learn the facial features and classify the emotions. The proposed model shows promising results, achieving 60% accuracy. Building on the model, a deeper analysis of the correlation between the emotions and the test takers grades is to be done, to understand which emotions are related to good grades. Applying LSTM neural networks to study temporal features in test videos to determine the emotion changes at different time intervals of the test and its application in the education domain is probable the future direction of research.

5. Acknowledgment

This work was supported by the College of International Cooperative Education, Harbin Engineering University. Special thanks go to the reviewers for the valuable feedback and Kaggle for making the fer2013 dataset available.

6. References

[1] Dana Lahat, Tulay Adal and Christian Jutten, “Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects,” Proceedings of the IEEE, Institute of Electrical and Electronics Engineers, vol.103 (9), pp. 1-10, 2015.

[2] Yousef Atoum, Liping Chen, Alex X. Liu, Stephen D. H. Hsu, Xiaoming Liu, “Automated Online Exam Proctoring” , IEEE Transactions on Multimedia, pp.1-4, 2017

[3] Ian Goodfellow, Joshua Bengio and Aaron Courville, “Deep Learning MIT Press”, http://www.deeplearningbook.org,pp.1-100, 2016

[4] Monika Dubey, Lokesh Singh, “Automatic Emotion Recognition Using Facial Expression: A Review,” International Research Journal of Engineering and Technology, vol. 3, pp.489 , 2016

[5] Kalani Ilmini and TGI Fernando,”Personality Traits Recognition using Machine Learning Algorithms and Image Processing Techniques”, Advances in Computer Science: an International Journal, Vol. 5, Issue 1, 2016

[6] Amogh Gudi, “Recognizing Semantic Features in Faces using Deep Learning”, arXiv:1512.00743v2 [cs.LG], pp.1-6,October 2016

[7] Nigel Bosch,Sidney K. D‟Mello,Ryan S. Baker, Jaclyn Ocumpaugh,Valerie Shute, Matthew Ventura,Lubin Wang and Weinan Zhao, “Detecting Student Emotions in Computer-Enabled Classrooms,” Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016

[8] Nabeela Altrabsheh, Mihaela Cocea and Sanaz Fallahkhair, “Predicting learning-related emotions from students textual classroom feedback via Twitter,” Proceedings of the 8th International Conference on Educational Data Mining, Madrid, June 2015

[9] Pooya Khorrami, Tom Le Paine, Kevin Brad, Charlie Dagli and Thomas S. Huang, “How deep neural networks can improve emotion recognition on video data”, MIT Lincoln Laboratory 2017

[10] Ikechukwu Ofodile, Kaustubh Kulkarni, Ciprian Adrian Corneanu, Sergio Escalera, Xavier Bar, Sylwia Hyniewska, Juri Allik, and n Anbarjafari, “Automatic Recognition of Deceptive Facial Expressions of Emotion”, Journal of IEEE transactions on affective computing,pp.1-10 , 2017

[11] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, “Deep Learning Review”, Macmillan Publishers Limited,pp.1-5 ,2015

[12] Veronica Perez Rosas, Rada Mihalcea, and Louis-Philippe Morency, “Multimodal Sentiment Analysis of Spanish Online Videos”, IEEE intelligent systems,pp.1-8(2013)

[13] Janez Zaletelj and Andrej Košir, Predicting students‟ attention in the classroom from Kinect facial and body features EURASIP Journal on Image and Video Processing (2017)