Sentence-Level Emotion Apprehension Through Facial Expression & Speech Verification Analysis

Md. Mohaimanul Haque
Department of CSE
American International University, Bangladesh
h.mohaimanul@gmail.com

Abu Fuzail polin
Department of CSE
American International University, Bangladesh
fuzail.polin@gmail.com

Souvik Das Dipta
Department of CSE
American International University, Bangladesh
souvik32100@gmail.com

Ashik Al Habib
Department of CSE
American International University, Bangladesh
ashikhabib32@gmail.com

Victor Stany Rozario
Assistant Professor(CS)
American International University, Bangladesh
stany@aiub.edu

***Abstract*—The importance of Emotional state apprehension is widely perceived in social interaction and social intelligence. Since the nineteenth century, this has been a popular research subject. In human-to-human communication, the understanding of facial expressions forms a communication carrier that offers vital data about the mental, emotional and even physical state of the persons in conversation. Inevitably user's emotional state plays an important role not only in human associations with other people but also in the way a user uses computers. As the emotional state of a person may determine consistency, task solving, and decision-making skills. Facial expression analysis, as used in this research, refers to computer systems that try to automatically predict user emotional state by analysing and identifying facial motions and facial feature changes from visual data. Though situations, body gesture, voice, individual diversity, and cultural influences, as well as facial arrangement and timing, all aid in interpretation. Facial expression analysis tools will be used in this research to analyse facial actions regardless of context, society, gender, and so on.**

***Index Terms*—Sentiment Recognition, Emotion Intensity, Soft-Max Classifier, Speech Recognition, Tokenization, Speech to Text, CNN, NLTK, Image Processing, Sentiment Prediction, Emotion Tagging, CMU-MOSEI.**

I. Introduction

A. Background of the study

This region includes a complete review of the strategies beforehand used to identify the emotion of a sentence. It can be shown that D.Das and S.Bandyopadhyay [4] used Conditional Random Fields (CRF) in their ML approach. The first stage is to use Affect List to generate an emotion for each word in a phrase, and the second stage is to discover the dominant emotion of each phrase using the influence points of each word. The first phase leverages word-level commentary using CRF, assessing information in the phrase that is already in the Affect lineup, and returning sentiment idiom categorization. The next step is to utilize these word-level preferences to derive the predominant component emotion from the frequency ratings of each word.

It has a high accuracy of 87.65%, which is constrained by the Affect list's Synonyms set (SynSet). This effort to identify words that do not exist in the said list failed.

To address the issue of sentence-level sentiment tagging, the authors of Paper [5] undertook an exploratory examination of alternative techniques. The first knowledge-based concept technique, WordNet-Affect Presence, was utilized to cope with comprehension alterations in a literature purely based on the presence of Affect vocabulary terms. In contrast, the Latent Semantic Analysis (LSA)[6] found the similarity between the provided text and each emotion independently in the following approach.

The third method makes use of equivalents mostly from WordNet SynSet. Finally, LSA emotive words, like WordNet Affect List, supports the previous set by mixing phrases from across all SynSets linked with a specific sentiment. This research also makes use of the Corpus linguistics-based method[7], which allows us to see how language is used in general as well as how it is historically employed. This strategy is more realistic, and it was also applied in our model.

The authors of paper [6] discuss a design for obtaining emotion tags using two systems: keyword spotting and lexical resemblance. Several techniques use the current lexical corpus to find words related to a specific emotion.

B. Problem Statement

The market for emotion-detection technology is currently worth $21.6 billion, and it is expected to more than double by 2024, reaching $56 billion. Businesses will buy systems to assist them in screening job applicants, analyzing ads for sentimental value, and testing criminal defendants for signs of fraud [19]. Suwa et al. adopted the facial expression identification system in 1978, manifesting an attempt to automatically understand facial expressions by monitoring the motion of

20 identified points on an image series [1]. The impact of the facial expression system on social interaction and social intelligence is widely recognized. Estimates of facial motion and identification of expression are all part of facial expression analysis.

Paul Ekman and Friesen carrying out more in-depth Darwin studies have concluded that there are six transcultural and prototypic emotional expressions [2]. These basic expressions are happiness, anger, sadness, surprise, disgust, and fear. Some authors consider the neutral face as a seventh expression [2]. It is known that facial expression changes through a set or subset of prototypic emotional expressions.

Given the enormous effect that Ekman's work had on the science of emotion, his views are prevalent in emotion-detection technology, influencing many algorithms, including those marketed by Microsoft, IBM, and Amazon.

C. Goals and Objective

Our primary aim is to create a model that can separate the fundamental emotions: angry, sad, surprised, disgusted, happy, neutral, and dread, and to achieve an accuracy higher than the standard 14.29 percent. We are now determined to examine our model's output in terms of the accuracy of each depicted class. Further development of the model is likely to result in more detailed classification with a more complex variance of the condition than lab condition images. We are determined to further develop the system that might include the frequency of human speech as well as gesture recognition.

II. Literature Review

In the Emotion detection research field, several domains contribute like machine learning, natural language, neuroscience, etc. In previous researches, facial expression, voice features, or textual data are used separately to classify emotions. Emotion can be classified into several static classifications like happiness, sadness, disgust, anger, fear, and surprise [20]. It can be further improved by combining the image, voice, and textual data. This combination of data gives further improved results.

A. Convolutional Neural Network:

To complete feature extraction and categorical classification, the emotion recognition challenge, like any other classification problem, involves using an algorithm. To distinguish an emotion, we must extract certain features from data and build a model that can categorize the input according to the feature. The procedure is outlined below:

- Data Processing:
  - The goal of data pre-processing is to standardize the data. The conventional method is to set the mean to 0 and then divide the data by the standard deviation.
- Feature Retrieval:
  - The typical method is to authenticate the identity and extract the Units from the face, and certain emotions might include combination of AUs code [21].
- Model Construction:
  - It will be either supervised or unsupervised. Support Vector Machine [14] is a prominent example of a supervised algorithm, whilst examples of unsupervised techniques include a Linear Discriminant Analysis.
- Result Generation:
  - The most common approach for producing a result is to identify which decision boundary has the shortest Euclidean distance from the inputs.

B. The issues with conventional method

Classification of facial expression can sometimes be difficult, surprisingly for humans also. Several studies have shown that people can identify different feelings in the same facial expression. It's still more difficult for AI. A variety of reasons make emotion recognition difficult. These elements can be classified as technical and psychological.

C. Technical challenges

Emotion detection has many issues detecting moving object in the video: detecting a target, uninterrupted detection, deficient or unforeseeable actions and many more. We have to overcome these challenges to obtain optimum model for our work. The emotion detection model examine facial features like eyes, noses, eyebrow position, chins, mouths and other features as actuation points. Occasionally, this detection intricate due to:

- The distance between features
  - There might be some functionality in the system that treats the average distance between landmarks as an epitome and compare them only within the range.
- Dimensions
  - The model may occasionally have difficulty spotting unusual characteristics such as slim or pale lips, small eyes, and so on.
- Tone
  - In rare circumstances, a solution may incorrectly identify a feature based on skin color.

Adopting a part-based approach that divides facial landmarks into distinct different sections based on the actual architecture of the face, according to [22][1]. The characteristics are then sent into the networks with the relevant labels by the model.

1) Data: When using machine learning or deep learning algorithms problems requires large amount of training data. Our data has videos with inconsistent frame rate, various angles and backgrounds, people of different genders and nationalities.

- Training Dataset
  - The training set (fer2013) consists of grayscale images of different layers with 48x48 dimensions. The data were automatically registered so that the faces are approximately centered and occupy a comparable amount of space in each image.
- Test Dataset

– For the Test Dataset we have used CMU-MOSEI multimodal Dataset. We had to pre-process the video file according to our needs.

2) Face occlusion and lighting issues: Occlusion happens when the target's posture is changed. This is frequent problem in motion detection in a video,[25] especially if the video is raw or unprepared. We can dissipate the problem by formalizing the system that is responsible for detecting the facial features in the video, identifying the features and deduce them to a 3D model of human face.We can further increase the efficiency of the model by introducing the illumination normalization algorithms.

3) Face occlusion and lighting issues: Most algorithms are designed to recognize high-intensity expressions. This might results in an inconsistent recognition of emotions when examining people from cultures with traditions of emotional suppression.

D. CNN

CNN models are composed of three distinct substructures: convolutional layers, pooling layers, and fully connected layers [23].

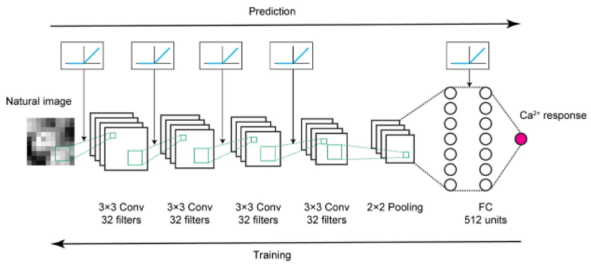

Fig 1: Convolutional Neural Network

A convolutional layer is made up of multiple feature maps. When constructing a new extracted features over the input, the local receptive field may be slithered [24]. When doing convolution, several data variants such as pictures and text are employed. Individual neurons in the layers may be linked to nodes in the receptive field rather than to all nodes or neurons in the previous layer.

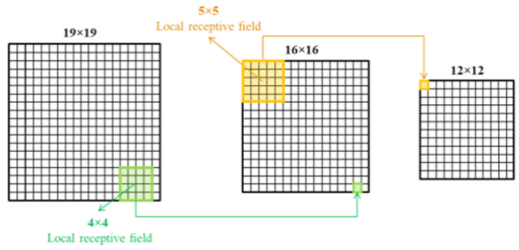

Fig 2: Local receptive field

To simplify the information, layer pooling is launched. These are commonly referred to as the subsampling layers. Pooling can be used to solve a variety of equations, including harmonic average, geometric average, and maximum pooling. The most common pooling procedures are max-pooling and average-pooling. These layers are mandatory to decrease the computational time.

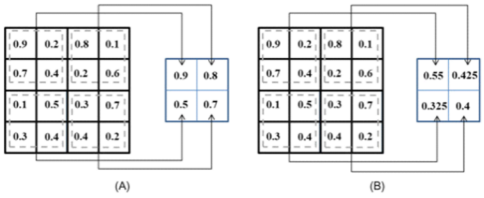

Fig 3: Max-pooling

Fully linked layers are the final layers of the CNN model, following the convolution and pooling layers. This layer allows the feature vector to be used as input data for certain machine learning problems such as prediction and classification[23]. SoftMax Classifier is another name for the final layer of completely connected layers.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$$

$$P(y = j \mid \mathbf{x}) = \frac{e^{\mathbf{x}^\top \mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^\top \mathbf{w}_k}}$$

Fig 4: SoftMax function

The max-pooling layer samples the extracted features down, reducing computation and guiding succeeding layers to focus on more significant details. In addition to the convolutional block, the entirely linked layers can serve as a classifier. Each unit of the layer contains the weight matrix, and the output is changed and activated to become the input of the future layers of units using the linear transformation and activation function [9].

Besides that, the last layer is typically a SoftMax classifier, with the conclusion being the one that has the maximum probability.

classI Multinomial(i)

label = arg max.

E. Speech Process

There are two types of data for computer- structured and unstructured. The structured data is highly organized and structured for processing and analysis within databases or spread sheets. Computers handle structured data very well. However, dealing with unstructured data or data that does not have a pre-defined structure or format, such as human language, is a challenge for them.

1) Stop words: Words that are super common and doesn't carry that much of a meaning, they just connect the important words of a sentence are called stop-words. For example, in English "a", "the", "are", "is" etc. are very common in pretty much every English sentence. So, the prescience is that by

removing these words, one can focus on the words that carry more prominence in a sentence or carry more information about the overall corpus. Different languages have different stop words. The Bengali language is a grammatically complex language. It has stop-words that needs to be removed for extracting meaning from a corpus. This technique commonly used in topic extraction, keyword searching, NLP classification tasks and so on.

2) Tokenization: Tokenization is the method of splitting a text stream into symbols of words phrases and other significant items called tokens of symbols may be individual phrases of words or even whole phrases. Tokens can be phrases, individual words or even entire sentences. Any characters that are not words like punctuation marks, are discarded in the process of tokenization. This tokenized word becomes input for things such as parsing and text mining.

To recognize the context, we used the open source python library Text2Emotion. It was developed with the goal of identifying acceptable attitudes in textual information. As an output, it constructs a lexicon that identifies circumstances into five basic emotion categories. such as **Happy, Angry, Sad, Surprise**, and **Fear**.

III. Methodology

A. Workflow

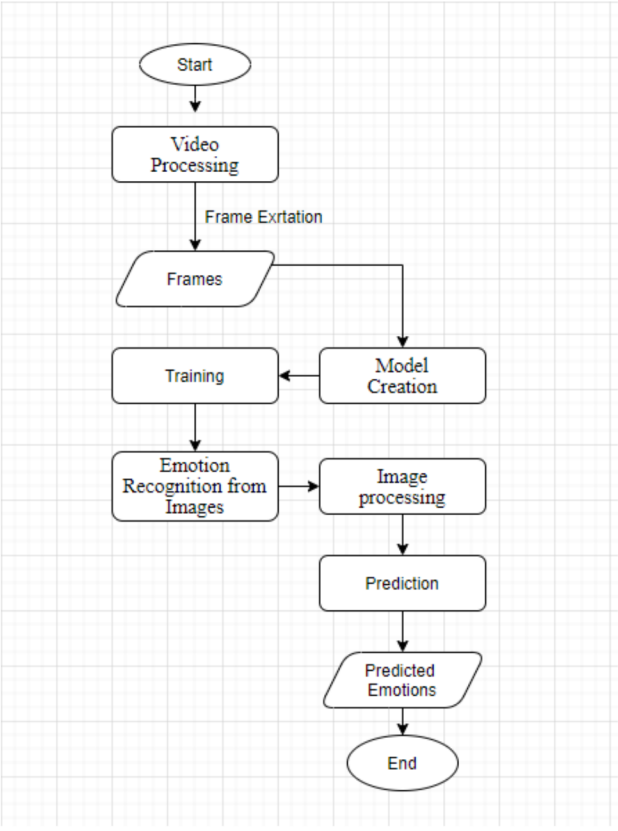

Fig 5: Work Flow (Frame recognition)

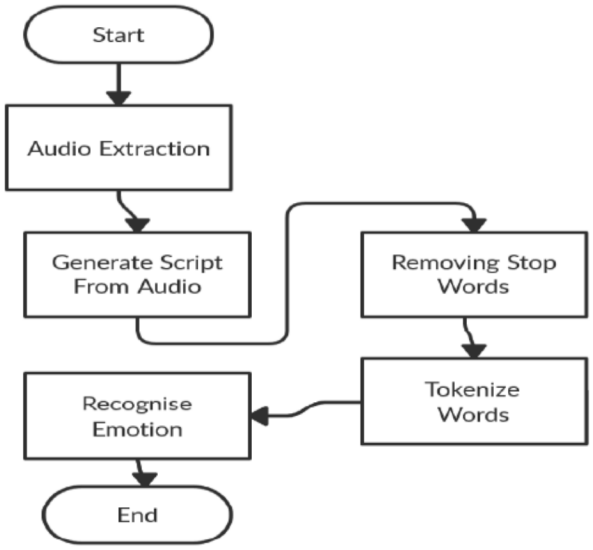

Fig 6: Work Flow (speech Recognition)

B. Video Processing

1) Video to Frame: First, we extract frame from the mp4 video which is located in Google drive. For extracting frame from video, we imported scikit-image, OpenCV-python library. First challenge is to have consistent data. Main rules in the dataset creation were:

- light conditions: records during the day
- video quality at least 640 width or above
- removed any cover with titles
- Face must be in the middle

Then we make each frame grayscale and resize it to 640. After that we stored every frame in Google Drive.

```
count = 0
success = True
vidcap = cv2.VideoCapture(videofile)
while success:
    if (count%one_frame_each == 0):# checks frame number and keeps one_frame_ea
ch
        success,image = vidcap.read() # reads next frame
        image_gray = rgb2gray(image)  # grayscale image
        if image.shape[1]>640:     # if image width > 640, resize it
          tmp = resize(image_gray, (math.floor(640 / image_gray.shap
e[1] * image_gray.shape[0]), 640),mode='constant')
        plt.imsave("%s/%s%d.png" % (OUTPUT_FRAMES_PATH,frame_name, c
ount), tmp, cmap= plt.cm.gray) # saves images to frame folder
        print ('*', end="")
        if count>=100:   #limits frames to 100
          break
    else:
        success,image = vidcap.read()# reads next frame
    count += 1
```

Fig 7: Frame Extraction from Video

2) Video to Audio: Initially, we have extracted audio from Video using MoviePy. MoviePy can edit all the several popular audio and video forms, including GIF, and runs on Windows/Mac/Linux, with Python 2.7+ and 3.

3) Reducing Noise: The default feature given by speech recognition which is (adjust for ambient noise) was used to reduce the noise of the audio.

4) Audio to Text: We have used and compared multiple voice recognition technology to achieve the best outcome. The technologies include google Text to speech, Microsoft Bing Speech, Wit, SoundHound for generating the script. On a given audio file, Google, Wit had shown similar results which are about 45% of the whole audio was extracted correctly where Bing showed comparatively better output about 61% of the whole transcript correctly. But finally, Soundhound was our API of choice for recognizing speech which is 80

C. Model Creation

For creating the model we use the Sequential API from keras library. Our Model contains four Convolutional Layers, two dense layer and one hidden layer.

1) Convolutional Layers: First Convolutional layer has 64 filters, size is 3x3, output matrix and input matrix are same. Second layer contains 128 filter, third layer contains 256 filters and last layer contains 512 filters. Other parameters were same.

Model is given bellow:

```
def createModel():
    model = Sequential()

    model.add(Conv2D(64, (3, 3), padding='same', input_shape=(48,48,1)))
    model.add(BatchNormalization())
    model.add(Activation('relu'))
    model.add(MaxPooling2D(pool_size=(2, 2), strides=None, padding='same'))
    model.add(Dropout(0.25))

    model.add(Conv2D(128, (3, 3), padding='same'))
    model.add(BatchNormalization())
    model.add(Activation('relu'))
    model.add(MaxPooling2D(pool_size=(2, 2), strides=None, padding='same'))
    model.add(Dropout(0.25))

    model.add(Conv2D(256, (3, 3), padding='same'))
    model.add(BatchNormalization())
    model.add(Activation('relu'))
    model.add(MaxPooling2D(pool_size=(2, 2), strides=None, padding='same'))
    model.add(Dropout(0.25))

    model.add(Conv2D(512, (3, 3), padding='same'))
    model.add(BatchNormalization())
    model.add(Activation('relu'))
    model.add(MaxPooling2D(pool_size=(2, 2), strides=None, padding='same'))
    model.add(Dropout(0.25))

    model.add(Flatten())

    model.add(Dense(512))
    model.add(BatchNormalization())
    model.add(Activation('relu'))
    model.add(Dropout(0.25))

    model.add(Dense(256))
    model.add(BatchNormalization())
    model.add(Activation('relu'))
    model.add(Dropout(0.25))

    model.add(Dense(7))
    model.add(Activation('softmax'))

    return model
model = createModel()
```

Fig 8: Model

D. Training

In the instance of Recognition from Video Frames, the aim is to classify each face into one of seven categories depending on the emotion expressed in the facial expression (0–6). (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral).The educational set contains over 28,000 cases, and the test set contains 3,500+ examples.

In case of Audio Recognition, the task is to categorize each word based on the meaning into one of five categories on the scale 0 to 1. Which are: Angry, fear, Happy, Sad and Surprise.

We train the model for 30 epoch and batch size of 8. Which game the model to have 71.5% accuracy.

E. Emotion Recognition from Images

For recognizing the images that were saved during video to frame conversion need to have some modification to fit into the model.

1) Image Processing: Firstly we need to detect the faces from the images. For face detection we use python keras.preprocessing library. Then the images were converter to have 48x48 shape.

2) Prediction: We used model.predict() function to predict the Emotion of the processed images taken form the videos. The predict() function takes an array of one or more data instances and enables us to predict the labels of the data values on the basis of the trained model. Then we take the max index of the predicted data and select the emotions.

F. Emotion Recognition from Audio

1) Removing Stop words: Words that are super common and doesn't carry that much of a meaning, they just connect the important words of a sentence are called stop-words. These should be removed too optimize and reduce valuable processing time. Here we have user NLTK (Naturak Language Toolkit) to Remove stop words from the Extracted Context.

Thus, the prescience is that by removing these words, one can focus on the words that carry more prominence in a sentence or carry more information about the overall corpus.

2) Prediction: We have used Text2Emotion python package to recognize the Emotion from the processed text extracted from the given video. Text2Emotion is a python package which automatically process texts, tokenize and extracts the emotions (Angry, fear, Happy, Sad and Surprise)

IV. Result & Discussion

The Video we used, was compiled into frames of every second and text was extracted from video

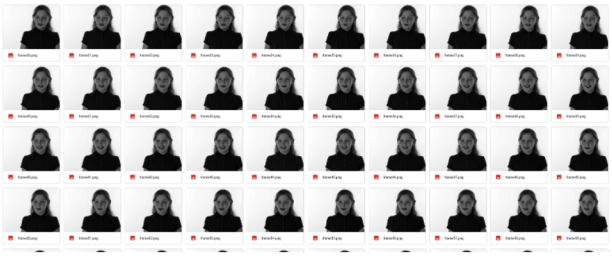

Fig 9: Frames taken from the video

100 Frames were taken throughout the whole video. Then The frames were resized into 48x48 grayscale image which was required according to our test Dataset. Using SoundHound Speech recognition Api support the recognized audio file was later transcript into a text as,

"what's the best thing about a gypsie on her period when you finger her you get here palm red for free biggest slut in history miss pac-man for twenty-five cents that is a pap smear called a pap smear because girls wouldn't do if it was called scrape the short side and while what do you call a cheap circumcision difference between a walrus and a lesbian."

After Removing stop words from the raw transcript it changed significantly into,

" what's best thing gypsie period finger get palm red free biggest slut history miss pac-man twenty five cents pap smear called pap smear girls called scrape short side call cheap circumcision difference walrus lesbian "

Thus, our text is ready for processing. Tokenization was automatically done through the process of Text2Emotion as its own feature.

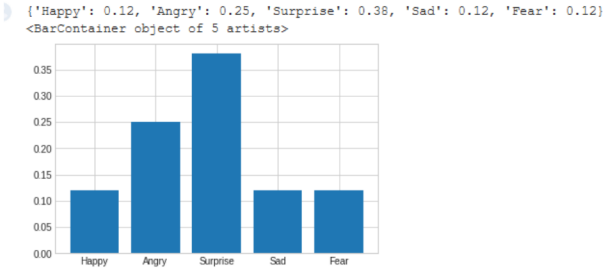

Fig 10: Results extracted from the video

The Above image shows the Result Found from the extracted context from the video. On the Other hand, from the video recognition we get,

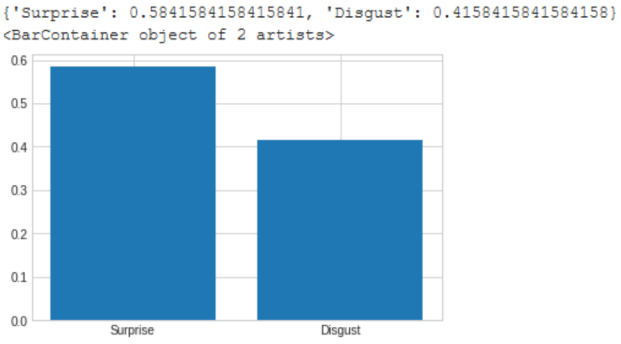

Fig 11: Video Recognition

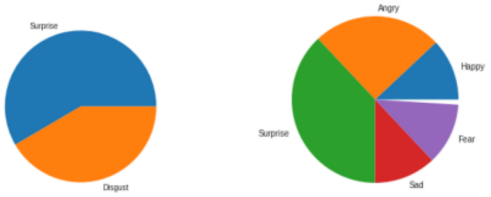

Fig 12: Visual representation Merged Result 2

Then we Merged Both result Dictionaries into one final dictionary given priority to the image recognized result. Which gives us a final picture of the recognition.

'Happy':	0.12
'Angry':	0.25
'Surprise':	0.5841584158415841
'Sad'	0.12
'Fear'	0.12
'Disgust'	0.4158415841584158

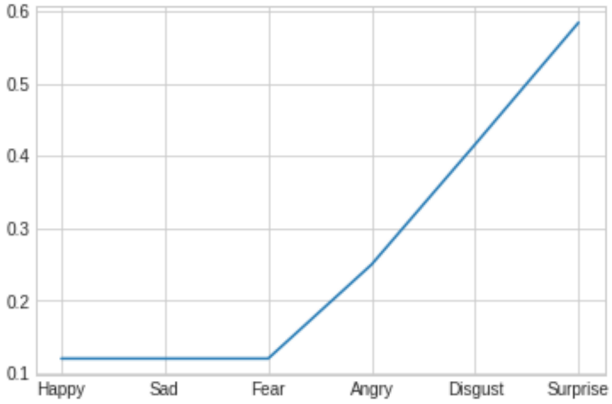

Fig 11:Visual representation Merged Result 1

V. Future Work

Few tests have been temporarily suspended due to a lack of appropriate resources (i.e. the experiments with real data are usually very time consuming, sometimes requiring even days to finish a single run). Future research will concentrate on a more in-depth look into specific processes, as well as plain curiosity.

There are certain adjustments that can be made to make the situation even better. Rather of using pre-existing data for training, a custom-made dateset is used. To further improve the accuracy of our model, gesture detection can be included. The most significant component of our effort, however, is to reduce the cost of applying the model to real-world issues.

References

[1] Andrea C. Samson, Sylvia D. Kreibig, B. Soderstrom, A. Ayanna Wad, "Eliciting positive, negative and mixed emotional states: A film library for affective scientists.", Cognition and Emotion, vol. 30, no. 5, 2015, pp. 827–856. https://doi.org/10.1080/02699931.2015.1031089

[2] Beatrice de Gelder ,"Why bodies? Twelve reasons for including bodily expressions in affective neuroscience.", Phil. Trans. R. Soc. B (2009) 364, 3475–3484. https://doi.org/10.1098/rstb.2009.0190

[3] D. Das, S. Bandyopadhyay : Sentence level emotion tagging. In: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–6. IEEE (2009). https://doi.org/10.1109/acii.2009.5349598

[4] Tegar S. Utomo, R. Sarno and Suhariyanto,"Emotion Label from ANEW dataset for Searching Best Definition from WordNet", International Seminar on Application for Technology of Information and Communication (2018) 249-252. https://doi.org/10.1109/isemantic.2018.8549769

[5] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." Psychological review, vol. 104, no. 2, p. 211, 1997. https://doi.org/10.1037/0033-295x.104.2.211

[6] Eissa M.Alshari and A. Azman, "Effective Method for Sentiment Lexical Dictionary Enrichment based on Word2Vec for Sentiment Analysis", Fourth Int. Conf. on Information Retrieval and Knowledge Management, 2018 https://doi.org/10.1109/infrkm.2018.8464775

[7] Asghar, Muhammad Zubair, et al. "Sentence-Level Emotion Detection Framework Using Rule-Based Classification." Cognitive Computation, vol. 9, no. 6, 2017, pp. 868–894., doi:10.1007/s12559-017-9503-3. https://doi.org/10.1007/s12559-017-9503-3

[8] Samson, Andrea C., et al. "Eliciting Positive, Negative and Mixed Emotional States: A Film Library for Affective Scientists." Cognition and Emotion, vol. 30, no. 5, 2015, pp. 827–856., doi:10.1080/02699931.2015.1031089. Strapparava C., Mihalcea R.: SemEval-2007 task 14: affective text. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), pp. 70–74 (2007) https://doi.org/10.3115/1621474.1621487

[9] [used] R. Dong, O. Peng, X. Li and X. Guan, "CNN-SVM with Embedded Recurrent Structure for Social Emotion Prediction," 2018 Chinese Au- tomation Congress (CAC), 2018, pp. 3024-3029 https://doi.org/10.1109/cac.2018.8623318

[10] Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In:Proceedings of the 2008 ACM Symposium on Applied Computing, pp. 1556–1560 (2008). https://doi.org/10.1145/1363686.1364052

[11] Shaheen, S., El-Hajj, W., Hajj, H., Elbassuoni, S.: Emotion recognition from text based on automatically generated rules. In: IEEE International Conference on Data Mining Workshop, pp. 383–392 (2014) https://doi.org/10.1109/icdmw.2014.80

[12] M. S. Ozerdem and H. Polat, "Emotion recognition based on EEG features in movie clips with channel selection," Brain Inf., vol. 4, no. 4, pp. 241–252, 2017. https://doi.org/10.1007/s40708-017-0069-3

[13] Hongli Zhang , Alireza Jolfaei , and Mamoun Alazab, "A Face Emotion Recognition Method Using Convolutional Neural Network and Image Edge Computing," IEEE ACCESS, Nov, 2019. 2949741 https://doi.org/10.1109/access.2019.2949741

[14] H. Ma and T. Celik, "FER-Net: Facial expression recognition using densely connected convolutional network," Electron. Lett., vol. 55, no. 4, pp. 184–186, Feb. 2019. https://doi.org/10.1049/el.2018.7871

[15] A. V. Savchenko, "Deep neural networks and maximum likelihood search for approximate nearest neighbor in video-based image recognition," Opt. Memory Neural Netw., vol. 26, no. 2, pp. 129–136, Apr. 2017. https://doi.org/10.3103/s1060992x17020102

[16] Salah, Albert Ali. Multimodal Behavior Analysis in the Wild Video-based emotion recognition in the wild. , (2019), 369-386 https://doi.org/10.1016/b978-0-12-814601-9.00031-6