International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-12, October 2019

Multimodal Decision-level Group Sentiment Prediction of Students in Classrooms

Archana Sharma, Vibhakar Mansotra

Abstract*: Sentiment analysis can be used to study an individual or a group's emotions and attitudes towards other people and entities like products, services, or social events. With the advancements in the field of deep learning, the enormity of available information on internet, chiefly on social media, combined with powerful computing machines, it's just a matter of time before artificial intelligence (AI) systems make their presence in every aspect of human life, making our lives more introspective. In this paper, we propose to implement a multimodal sentiment prediction system that can analyze the emotions predicted from different modal sources such as video, audio and text and integrate them to recognize the group emotions of the students in a classroom. Our experimental setup involves a digital video camera with microphones to capture the live video and audio feeds of the students during a lecture. The students are advised to provide their digital feedback on the lecture as 'tweets' on their twitter account addressed to the lecturer's official twitter account. The audio and video frames are separated from the live streaming video using tools such as lame and ffmpeg. A twitter API was used to access and extract messages from twitter platform. The audio and video features are extracted using Mel-Frequency Cepstral Co-efficients (MFCC) and Haar Cascades classifier respectively. The extracted features are then passed to the Convolutional Neural Network (CNN) model trained on the FER2013 facial images database to generate the feature vector for classification of video-based emotions. A Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM), trained on speech emotion corpus database was used to train on the audio features. A lexicon-based approach with senti-word dictionary and learning based approach with custom dataset trained by Support Vector Machines (SVM) was used in the twitter-texts based approach. A decision-level fusion algorithm was applied on these three different modal schemes to integrate the classification results and deduce the overall group emotions of the students. The use-case of this proposed system will be in student emotion recognition, employee performance feedback, monitoring or surveillance-based systems. The implemented system framework was tested in a classroom environment during a live lecture and the predicted emotions demonstrated the classification accuracy of our approach.

***Keywords:* multimodal, sentiments, deep learning, convolutional neural networks, recurrent neural networks, support vector machines, classrooms**

Revised Manuscript Received on October 30, 2019.

***** Correspondence Author

Dr. Archana Sharma*, Department of Computer Science, Government M.A.M College, Cluster University of Jammu, Jammu, India. Email: archana.35188@yahoo.com

Dr. Vibhakar Mansotra, Department of Computer Science and IT, University of Jammu, Jammu, India. Email: vibhakar20@ yahoo.co.in

I. INTRODUCTION

The study and analysis of sentiments is gaining broad popularity for its invaluable insights into people's behavior and emotions based on the voluminous data available in today's world. It can be used to detect the polarity in the analyzed data, i.e. positive or negative sentiments. This has become the primary approach used in building recommendation engines for movies, online products etc., reviewer systems for customer feedback on products, and used in predictive analytics to estimate the finance, budget etc. This understanding of a human's attitude or opinion regarding a specific substance or thing has uses in a variety of applications.

To this day, most of the research in this field have been actively carried out in the natural language processing like analysis of textual data. However, with the increased popularity of social media platforms [1] , people have started to express their views on varied topics like politics, governance, economics, technology etc., through videos in platforms like YouTube, Facebook, Twitter etc., and images in platforms like Instagram, Facebook, Flickr etc., and audio using podcasts or WhatsApp audio shares. Hence, it has become the need of the hour to mine these opinions and beliefs to identify the underlying sentiments from diverse modalities like audio, video and text. This growth in the social media platforms has pushed the traditional text-based sentiment analysis [2] to develop more complex learning models to accumulate information from diverse forms – videos, images, audio, text etc.

There have been numerous techniques used in the past researches for analyzing the sentiments using automated machines. Neural networks stand out of the lot as it has been modeled around the human brain and the neurons are capable of learning from new data and correlate the learned information to the past experiences to subjectively decode the data. Deep learning systems have advanced software design, adaptable learning procedures, and access to computing power and training data. It has already found its place in research areas like computer vision, speech recognition and Natural Language Processing (NLP) [3] and diverse field areas like object detection, robot navigation, visual classification etc. Deep learning has been embraced today by the researchers working on sentiment analysis for its architectural sophistication, learning prowess and the ability to work with both supervised and unsupervised methods.

As the name suggests, *"Multimodal Sentiment Analysis"*, [4] inspects other means for analyzing data, particularly the audio and video data. This combination of facial expressions and vocal signals in conjunction with textual data can positively enrich the learning process of the machine learning models to better categorize the mental states of the people whose opinions are analyzed. This can help the system better understand the sentiments from the behavioral clues in the visual and vocal signals. Deep learning models such as "*Convolutional Neural Networks*" (CNN) [5], "*Deep Belief Networks*" (DBF) [5], *"Recurrent Neural Networks"*(RNN) [5] etc., have been deployed in text generation, feature representation, sentence classification, vector representation, object detection, speech estimation and face recognition and have proved to be extremely beneficial. In this paper, we propose to use deep learning models such as RNN to analyze the speech signals and CNN to study the facial features to automatically detect the underlying sentiments experienced by the person.

Multimodal Emotion Recognition [4] presents a challenge in the integration of the audio, text and visual features to effectively determine the analyzed sentiments. Since multimodal refers to the fusion of all single modalities to comprehensively represent a single entity, fusion techniques such as feature-level and decision-level can be used to reliably represent the recognized sentiments from the multimodal information gathered from different streams. Feature-level gathers all the features available from each procedure and creates a single feature vector that is fed to the classification algorithm. But it may encounter difficulties if our approach requires us to integrate heterogeneous features. Decision-level uses the extracted features from each approach and feds them individually to its own classification algorithm. The overall classification results are achieved by fusing together the results from each modality into a single decision vector. Since this does not attempt to fuse the diverse features from different sources like video, audio and text, the classification algorithm can be selected to perfectly suit the corresponding modality. For the same reason, we have decided to use decision-level integration in our paper to make it work better for every modal approach like video, text and audio.

The focus of our paper lies in building a multimodal sentiment recognition [4] system that can be used to analyze the emotions of the students in classrooms. This is essential in the present scenario, with the advent of digital classrooms [6], the interaction between the student and the instructor has decreased considerably. Also, with smartphones and social media, there are innumerable distractions to the students, letting them to lose focus and their attention span has reduced significantly. This initiates the necessity to develop a multimodal system that can capture their emotions in real-time and associate it with a feeling that clearly represents their mood in the classroom. Our system uses a video camera to capture the faces of the students along with their speech audio and the frames are extracted to split the audio and video for processing and classification. We have proposed to capture the student emotions conveyed via Twitter, to the department twitter handle. The classified emotions from these multimodal systems can be integrated using decision-level features and their precise sentiments can be represented for intellectual analysis. This implementation can be hugely beneficial to both the students and the instructors to learn from it and help them to modify their approach to learning and teaching respectively.

II. RELATED WORKS

There are only a few multimodal researches works other than survey that combine the three approaches such as visual-act, speech-act and text modality. Here, we have considered a few literature studies based on visual, speech, twitter-based and multimodal emotion recognition systems. Each approach uses respective methodologies and datasets to identify emotions.

A. Visual Emotion Recognition

In their research work, Ugur Avyaz et al. [7], proposed a methodology that detects the emotional states of students in a classroom. A customized educational dataset was used for the determination of emotional states which contains face images of twelve undergraduate students. Viola-Jones algorithm was used to detect the faces from a real-time image of classroom. Four machine learning algorithms such as K-nearest neighbors (kNN) [8], Random Forest [8], Decision Tree [8] and Support Vector Machine (SVM) [8] was used to classify the emotional states and compare the results one another. On those, SVM has got the best result which is more than 97% accuracy, whereas decision tree regression method got poor results.

Yang et al. [9] designed an emotion recognition model on facial expressions. This implementation helps educator to understand the behavior or emotions of students in virtual learning environments. This method used the JAFFE database [10] that performed well in recognizing emotions. Haar cascade classifier was used to detect and extract the facial characteristics. Sobel edge detection [11] technique was used to filter and edge detection of the input images for feature extraction. Artificial neural networks were used to train and classify the emotions of a student. This algorithm achieved moderate results due to some problems in virtual learning environments such as lack of internet speed, quality of webcam.

In another prior study, Chao Ma et al. [12], proposed an emotion recognition system used in online education environments with deep learning algorithms. FER2013 database was used for training. Haar feature real-time detection algorithm based on Adaboost [13] was used to detect and extract the faces from the input image of the students. A convolutional neural network algorithm was applied on FER2013 dataset to train and predict the emotions. To improve robustness and prevent overfitting, Gaussian noise [14] was added to the input images. An emotion weight score ranging from [-0.9 to 0.9] in five different intervals was calibrated for each emotion. This method helps for distance education students by providing feedback of their emotions to the lecturer for delivering quality online education.

B. Speech Emotion Recognition

Ling Cen et al. [15] implemented an emotion recognition system by using verbal communication.

Verbal communication plays a vital role in reflecting speakers' response which are expressed in different emotions. A customized GUI interface was designed to understand the technical information of speech-act modality recognition system. The speaker's audio was taken in three different ways which are, live audio recording, single recorded file and multiple recorded files. A voice activity detection (VAD) [16] algorithm was used to detect the human speech and avoid the non-speech sections. A finite impulse filter (FIR) [17] was used to emphasize important frequency pulses in the speech signal. In order to train the model, audio signal should be converted to features. In this study, three types of cepstral features were used, i.e., PLP cepstral coefficients, MFCC and LPCC. Machine learning based support vector machines was used to classify the emotions. This study achieved better results, but the results were poor in the case of negative emotions. In their work, George Trigeorgis et al. [18], proposed a speech emotion recognition system using deep learning architecture. This study used a popular approach to classify the emotions. A popular and natural emotions RECOLA database [19] was used for training. FIR filter was used to reduce the noise as discussed by Ling Cen et al. [15]. A raw waveform with a six second interval was segmented for training instead of extracting the acoustic features, because the complexity of the feature extraction is quite low. A CNN model was associated with LSTM architecture was used to classify the emotions. This model significantly achieved better performance.

Seyedmahdad Mirsamadi et al. [20] proposed a speech emotion recognition system using recurrent neural networks. This study used a new speech corpus database named as Interactive Emotional Dyadic Motion Capture (IEMOCAP) [21] which is a multi- speaker database. This database has achieved a moderate performance using recurrent neural network during the speakers are in dyadic interaction. A short-time frame level acoustic features and sequential combination of those features in utterance level representation with the help of Low-level Descriptors (LLD) [22] was used for training the recurrent neural network. This study trained with RNN and a joint learning (RNN-LSTM). But, unfortunately RNN model achieved better performance results than the joint learning method.

C. Textual Emotion Recognition

Hassan Saif et al. [23] had implemented a sentimental analysis mechanism from twitter posts by using lexicon-based approach. SentiCircles, a lexicon- based approach was proposed in this paper. It is mainly based on the contextual semantics which represents the sentiment orientation of words. The proposed method was evaluated with three different datasets which are Obama-McCain Debate (OMD), Health Care Reform (HCR) and Stanford Sentiment Gold Standard (STS-Gold). A data acquisition and data annotation were used to extract the named entities and labelled with appropriate emotions from a dataset collected dataset. This approach used the MPQA subjectivity method and SentiWordNet lexicons to predict the sentiment from a text. This study has proved to be the best in lexicon-based approaches.

Bac Le et al. [24] proposed a sentiment analysis system from twitter data. As we know that, there are few approaches to do sentiment analysis from text and we have already seen the earlier study of Hassan Saif [23] using lexicon-based approach. A learning-based approach was used in this paper. They have done experiment on three different datasets which are Alchemy API, Zemanta and Open Calais. Naïve Bayes and SVM classifier were used to train these three datasets and classify the emotion individually. Out of all those, Alchemy API achieved a better performance with two classifiers.

Orestes Appel et al. [25] presented a hybrid approach for sentiment analysis. Hybrid approach is a combination of lexicon and learning based approaches. A SentiWordNet dictionary was used in lexicon-based approach and twitter movie review dataset was used in learning-based approach. A Hybrid Standard Classification (HSC) was used to classify the emotions for lexicon-based approach and Hybrid Advances Classifier was used to classify the emotions for learning based approach. A sentiment polarity will be assigned to these approaches and identify the actual emotion. This study achieved better results of about 75% accuracy on a movie review dataset.

D. Multimodal Emotion Recognition

Ya Li et al. [26] proposed an audio- visual emotion recognition system. The goal of this study was to apply machine learning methods on multimedia data for multimodal emotion recognition. A Chinese Natural Audio-Visual Emotion Database (CHEAVD) was used for training. This dataset has video data, extracted audio data and corresponding emotion label. Viola Jones algorithm was used to detect the faces in the video frames. Local binary pattern was used to reduce the feature vector and extract the features from the facial expressions. OpenSmile toolkit was used to extract the audio features. Different machine learning methods were applied on audio-visual features, but random forest-based classifier generated around 40-50% accuracy which is still below par. Soujanya Poria et al. [27] proposed a multimodal emotion recognition system which has audio, visual and text-based classifications. In this study, a multimodal opinion utterance database (MOUD) was used for textual data training and IEMOCAP dataset [21] was used for multimedia data training. MOUD database has a collection of multiple social media videos of either reviews or recommendations. IEMOCAP has twelve hours video data contains spoken utterances and labelled emotions. CNN was used to train the visual dataset, RNN based deep learning technique was used to train the speech data and Multiple Kernel Learning (MKL) [28] was used to train the textual data. This proposed model integrated all these approaches and classified an accurate emotion. This study has achieved more than 75% accuracy and resulted in a better performance for multi-view emotions.

III. PROPOSED SYSTEM FRAMEWORK

The implementation methodology has been designed in a high configuration GPU enabled computer with the deep learning framework, TensorFlow installed. We conducted some of our model implementations and their training in Google Collaboratory, which is a free cloud service provider with GPU and TPU. Earlier, we had implemented our emotion recognition systems using facial, speech and text [29] as individual works.

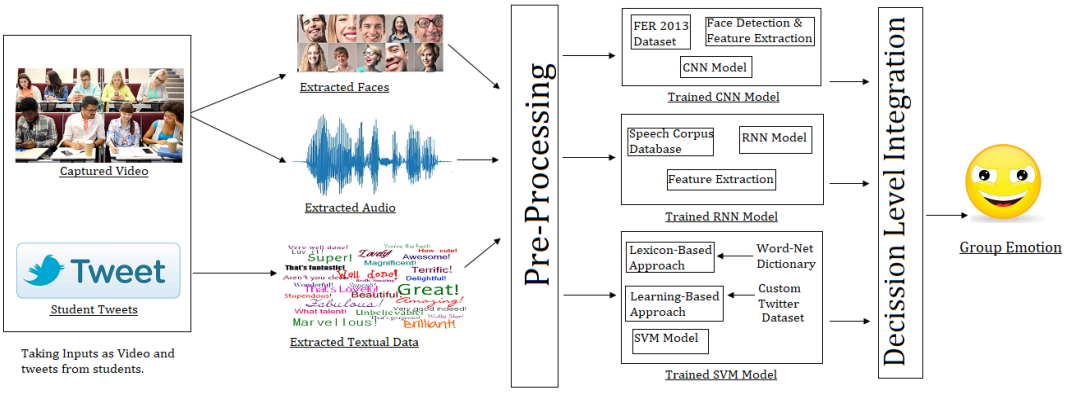

Fig. 1. Architecture of Proposed Multimodal Sentiment Recognition Framework

In this paper, we have integrated those approaches and emotion recognition methodologies and applied a decision-level integration technique to achieve the precise classification of student sentiments.

The model architecture of proposed framework is shown in Fig 1. We have discussed in detail about the preliminary preparations required to get the datasets ready to be trained by the deep learning models and predict the emotions as detailed in the blocks represented in Fig 1.

A. Dataset Preparation

Dataset is the foremost step for training the machine learning or deep learning models. Datasets act as an input to the model and splits up into train and test sets for initiating the learning process. Train data is used for training the model and test data is used for evaluating the efficiency of the model.

In this study, we have used the Facial Emotion Research (FER) database [30]. This database has 48x48 pixel size images of faces and it can predict seven different emotions such as anger, fear, sad, happy, neutral, disgust and surprise. Each image in this database is labelled with appropriate emotions as a numerical code ranging from 0-6. Each database has unique advantages such lighting, motion, angle and emotions. Due to consideration of all these constraints and reviewed few other facial databases such as CK+, MMI...etc., FER database is more suitable to our application and helped to achieve the promising results on identifying emotions of students in our previous research work.

In speech emotion recognition, a Speech emotional corpus database was used to train and evaluate the model. This database consists twelve hours of audio recorded with four male and five female speakers in a recording room with calm and quiet environment. It classifies six different types of emotions such as anger, sad, happy, neutral, surprise and fear. There are a few constraints needed to be taken care of while picking the dataset such as local accent, use-case, emotion classes, noiselessness. So, we have chosen this database by taking all these constraints into account.

In textual emotion recognition, we have manually constructed a twitter emotion dataset, large enough to prove to be efficient for learning, by collecting the twitter posts from the department handle or the lecturer's twitter account which contains the comments, feedbacks or opinions that may be expressive of the sentiments. We have gathered all the text and manually labelled with appropriate emotions and have split it into train and test data. This dataset had performed well in both training and evaluation.

B. Pre-processing

Pre-processing is the most crucial step in the training process as every database may have noise, inappropriate data or errors. It prepares the data for further processing. There are many things involved in preprocessing stage.

1) Facial Data Pre-processing

Face detection would be the first and foremost step in pre-processing. There are few algorithms to detect the face in an image. Viola-Jones algorithm was used in this study. This algorithm uses Haar basis filters to identify the facial features to detect the faces. Once the face is detected, it crops the facial images and sets a fixed common size of 48x48 pixel. If any grayscale images were present in the database, the whole database must be converted into grayscale images which can be represented as 8-bit data whereas color images are 24-bit. Once preprocessing is done, the resized images can be moved for further processing.

2) Speech Data Preprocessing

The preprocessing stage for speech modal analysis can improve the vocal features by using certain techniques such as noise reduction, silence removal, pre-emphasis filter, framing and windowing.

- Noise reduction phase delivers clean signal and removes corrupted signal. To do this, minimum mean-square error (MMSE) technique was used.
- Silence removal can be an efficient preprocessing scheme to eliminate background noise and unvoiced segments in the speech signal.
- Pre-emphasis filter is the process of amplifying high frequencies in the audio signal. It also allows balancing the frequency spectrum and avoiding numerical problems in Fourier transform computation.
- Framing allows to split the speech signal into short-term intervals or windows. Usually, it ranges from 20ms – 50ms window size. In our case, we have taken 25ms frame size.
- Windowing is the process of reducing spectral leakage or signal discontinuities. It can allow to improve the signal clarity. Hamming window is popular windowing algorithm which we have used in our study.

3) Textual Data Preprocessing

The techniques involved in processing and cleaning the text data is discussed below.

- Weighting scheme is the standard way to construct the feature vector. It identifies the important words in a database by checking the term frequency by making the use of the term frequency-inverse document frequency (TF_IDF) approach.
- Stemming algorithm allows to remove the stem features (suffix word) of a word. For e.g. 'playing' has many suffixes such as played, play. It will consider all these words as the same feature.
- Tokenization allows to construct a word vector named as bag-of-words by splitting the documents into words.

C. Feature Extraction

Feature extraction transforms original data into vectors of data which contains more discriminatory information. The feature extraction methods used in our study for each modal scheme is discussed below.

1) Facial Data Feature Extraction

In facial image data, face extraction identifies different features such as eyes, nose, forehead, mouth…etc. The CNN model was used to deduce the feature vector from the facial features. It can handle various face images with different poses, facial expressions and illumination. The CNN approach applied here can provide high computation and better performance.

2) Speech Data Feature Extraction

The raw audio files from the speech emotion dataset cannot be used directly for training the classification model. So, it is essential to extract those numerous discriminative features present in the raw audio files. To achieve it, we have used Mel-Frequency Cepstral Coefficients (MFCC) [31] algorithm to extract the specific features such as pitch, vocal tract rate and power. MFCC uses Discrete Fourier transform (DFT) to extract the features which can be used to recognize whether the voice is human or not.

3) Textual Data Feature Extraction

Feature extraction in textual data gives maximum support to machine learning model by computing positive and negative polarity of the sentences which provides most of the opinion. Some example features in the feature extraction stage are parts of speech tags, opinion words and phrases, positions of terms and negation. In textual kind of emotion representations, most of the work has been done by preprocessing and feature extraction.

D. Training

Machine learning or deep learning algorithms will be used for training and classifying the data. As we discussed in dataset preparation section, the input to the neural network algorithm must be training set.

1) CNN model for Facial Emotion Recognition

A three-layer convolution neural network model with transfer learning (VGG19) was used in this study. The CNN model creates convolution kernels by performing convolutions on the facial dataset by applying corresponding weights to the extracted features. The overview of proposed CNN model shown in Fig 2.

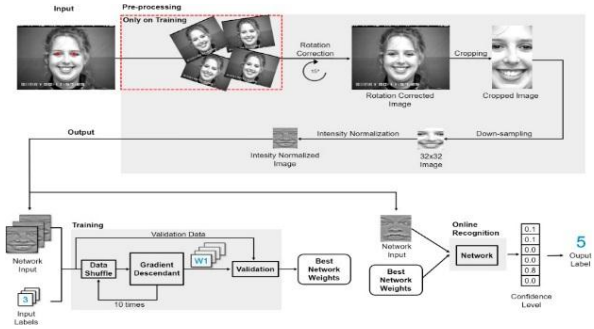

Fig. 2. Overview of the Proposed Convolutional neural network model [32].

CNN has three types of layers; input layer, hidden layer and output layer. The input layer is fed with 48x48 pixel size extracted features of training set. The hidden layer contains activated function and learned features from the input layer. The output layer consists of emotion classes learned from the hidden layer. Each convolution cyle involves specific layers namely convolution, pooling and normalization.

In order to implement the CNN model, we have used a TensorFlow [33], a popular open-source framework with Keras backend. Keras is a high-level neural network API that runs on TensorFlow framework. A VGG-19 neural network architecture, a pre-trained model was used to improve the performance and accuracy of our proposed method. To do that, a transfer learning approach was used in our study.

Pseudo Code : *Facial Emotion Recognition Training*

1. **Begin**
2. Assign N = number of training images
3. **For** i:=1 to N **do**
4. Obtain an i^{th} image, data = num_imges(i)
5. Obtain the length of data, l = length(data)
6. **For** j:=1 to l **do**
7. Obtain j^{th} face, i_j = FD(j)
8. Obtain features from i_j, F_j = FE(i_j)
9. **End for:**
10. Extracted features from all images of given training dataset $X_i = F_1 \,||\, F_2 \,||\, F_3 \;\ldots . F_i$
11. **End For**
12. Assign emotion label for each i^{th} image to Y_i.
13. Use all training features (X), and emotion labels (Y) to train the CNN model.

E. End

Note: *FD = Face Detection, FE = Feature Extraction*

2) RNN model for Speech Emotion Recognition

A recurrent neural network with Long-short term memory architecture was used. RNN uses sequential nature of information whereas inputs are independent of each other. As you can see in Fig 3, the parameters (U, V, W) are shared across all the steps in the network. In this case of LSTM, we need not produce output in each step, as the output layer will produce the probabilities of vectors at the end of the steps. As the RNN parameters are shared across all the steps, the gradient at each step of the output depends on both the current time step and previous time step which is termed as Back Propagation. LSTMs have great advantage of remembering output for long propagations and classify the emotions by probability scores.

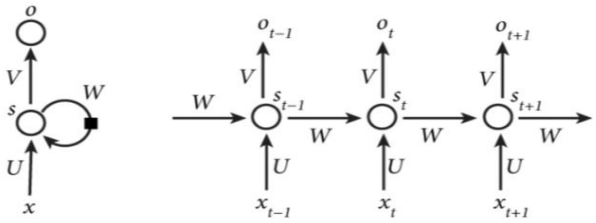

Fig. 3. A schema of Recurrent neural network

Pseudo Code : *Speech Emotion Recognition Training*

1. **Begin**
2. Assign S = length of the training audio sample
3. Assign t = 20sec
4. **For** i:=1 to S **do**
5. /* Splitting the audio sample into t intervals */
6. Obtain audio samples, T = split(S, 20)
7. Obtain the number of samples, N = length(T)
8. **End for**
9. **For** j:=1 to T **do**
10. **For** k:=1 to N **do**
11. Obtain k^{th} audio sample, p_k = PP(j)
12. Obtain features from p_k, F_k = FE(p_k)
13. **End for:**
14. Extracted features from all audio samples of given training dataset $X_i = F_1 \,||\, F_2 \,||\, F_3 \;\ldots . F_i$
15. **End For**
16. Assign emotion label for each i^{th} audio sample to Y_i.
17. Use all training features (X), and emotion labels (Y) to train the RNN-LSTM model.
18. **End**

Note: *PP = Pre-Processing Face FE = Feature Extraction*

3) SVM model for Textual Emotion Recognition

Support vector machines are widely used for classifying textual data. In this work, since the crucial steps were accomplished by preprocessing and feature extraction, there did not arise any need for a deep learning algorithm. SVM uses non-linear kernel functions that transforms input vector of features into high dimensional feature space. There are only three things needed to focus on support vector machines algorithm in textual emotion. First, a proper dataset which is labelled and split into train and test data. Second, vectorizing the data which means extraction of feature vectors. Third, build a support vector machine for training. Our SVM model presented the reasonable and satisfied results.

Pseudo Code : *Textual Emotion Recognition Training*

1. **Begin**
2. Assign n = number of tweets in dataset
3. Assign m = number of emotions to classify
4. Assign S = $[s_1, s_2, s_3 \ldots s_n]$ as an input of tweets vector
5. Assign E = $[e_1, e_2, e_3 \ldots e_n]$ as an output of emotion labels
6. **For** i:=1 to length(S) **do**
7. Obtain a i^{th} sentence, data = PP(i)
8. Obtain the features from i^{th} sentence, F_i = FE(i_i)
9. **End for:**
10. **For** k:=1 to F **do**
11. Apply a lexicon-based approach on Feature Vector of data F, L = LBA(F)
12. **End for**
13. Obtain an Input vector $X = F_1 \,||\, F_2 \,||\, F_3 \;\ldots . F_i$
14. Obtain an Output vector $Y = L_1 \,||\, L_2 \,||\, L_3 \;\ldots . L_i$
15. Use Textual Features X, Emotion labels Y to train a Support vector machine and classify the emotions.
16. **End**

Note: *PP = Preprocessing, FE = Feature Extraction LBA = Lexicon-based Approach*

E. Validation

The proposed method was experimentally tested in a classroom environment. Once the training was completed, a trained model was built using Keras in HDF5 format and saved to a local storage for each approach. A live video session of the lecture with the students live participation will be captured with a video camera assembly fixed in the classroom. A twitter API will gather live tweets posted by students at the end of every lecture session. FFMPEG is a python library that can extract the audio from video frames. The proposed system processes audio, video and text data by applying pre-processing techniques such as face detection, noise removal, text tokenization. The extracted features are given as an input to the respective deep learning models that are trained on their corresponding datasets and the sentiments are represented. The classified features from each modal approach representing the sentiments as a group emotion is generated by analyzing the student's emotion probability scores. Once all the emotions are identified from each approach, a decision level integration that uses combination associations and correlations to arrive at a decision takes place as shown in Fig 4 and the group emotions are predicted by considering the accuracy, precision rates and probability scores.

IV. Results And Discussion

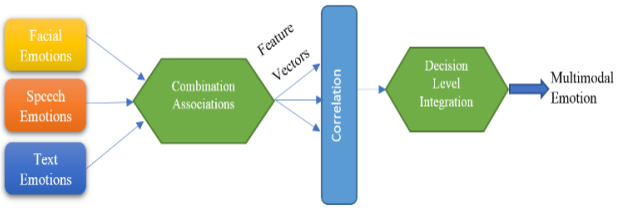

Fig. 4. Decision Level Integration

As discussed, we have implemented this proposed framework in a classroom environment where students can express their feelings, emotions and opinions via facial, speech and tweets. The experiment was arranged to work in a relatively noise free classroom environment. However, noise generated by mischievous students and other external noises had a severe impact on the speech modal. The facial expressions were captured better even in different head poses, angles or tilts. Overall, our system was able to achieve an accuracy of about 75%. We have tabulated the accuracy and precision of the framework, determine from the classification results as shown in Table -I.

TABLE – I: ACCURACY AND PRECISION VALUES OF PROPOSED FRAMEWORK

Modality	*Accuracy*	*Precision*
Facial	0.812	0.421
Speech	0.776	0.614
Text	0.831	0.376
Multi-modal	**0.762**	**0.525**

The confusion matrix representation for the decision level integration is as shown in below Fig 5. The numbers from 0-6 represents the emotions Anger, Fear, Sad, Neutral, Happy, Surprise, Disgust. The mean diagonal intersection of true and predicted labels in the confusion matrix correspond to the average accuracy of the decision-level framework.

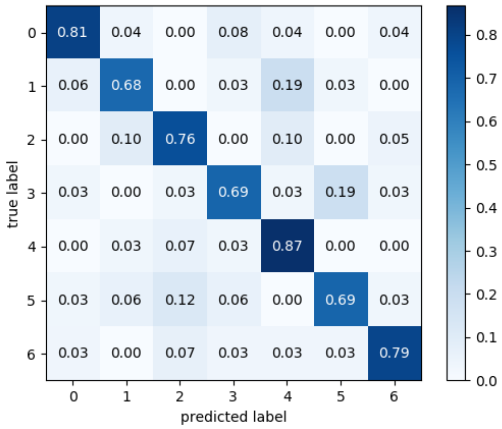

Fig. 5. Confusion Matrix

We have calculated the success rates of each emotion and plotted in a bar graph. The emotions "*Happy" and "Anger"* achieved more than 80% success rate i.e., those emotions are identified easily and the emotions "*Sad*" and "*Neutral*" had less success rates which are around 60%. The bar plot of success rates on each emotion as shown in Fig 6.

Fig. 6. Bar Plot showing Success Rates of Classified Emotions

V. CONCLUSION

This paper presents a multimodal framework for the analysis and classification of students' sentiments with the appropriate features extracted for the facial, speech and twitter text data. It also offers a simple, but effective technique to integrate the extracted features acquired from the three modal presentations. As discussed in the earlier sections, Haar cascades based facial features extraction was very helpful for emotion classification. MFCC based audio feature evocation was of immense assistance to isolate the speech waves among other noises and generate reliable set of features. We had selected the transfer learning approach in CNN for the classification of facial emotions, primarily, due to its ability to reduce the time spend in training the model, especially with the large FER2013 dataset. This in turn resulted in improved predictions and classification accuracy. For the classification of sentiments using speech, we have combined the RNN approach with LSTM which can be more efficient when working with large and non-linear datasets. The twitter text-based classification of sentiments was done using a hybrid approach with lexicon and SVM based feature learning that performed exceptionally well. Our decision to use decision-level fusing of the features worked well in integrating the sentiments from the different modal schemes into one valid feature representing the classified sentiments. To conclude with, we are satisfied with the results achieved and look at the possibility of developing multimodal sentiment recognition system that is language and culture independent.

REFERENCES

1. Asur, Sitaram, and B.A. Huberman. "Predicting the future with social media." In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 492-499. IEEE Computer Society, 2010.
2. B. Pang and L. Lee. "Opinion mining and sentiment analysis." *Foundations and Trends® in Information Retrieval* 2, no. 1–2 (2008): 1-135.
3. Collobert, Ronan, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. "Natural language processing (almost) from scratch." *Journal of machine learning research* 12, no. Aug (2011): 2493-2537.
4. S. Turol. (2018, August 22). "Multimodal sentiment analysis with TensorFlow"[Online].https://www.altoros.com/blog/multimodal-sentiment-analysis-with-tensorflow/.
5. Soleymani, Mohammad, D. Garcia, B. Jou, B. Schuller, S. Chang, and M. Pantic. "A survey of multimodal sentiment analysis." *Image and Vision Computing* 65 (2017): 3-14.
6. Dede, Christopher, and J. Richards, eds. *Digital teaching platforms: Customizing classroom learning for each student*. Teachers College Press, 2012.
7. U. Ayvaz, and H. Gürüler. "Real-time detection of students' emotional states in the classroom." In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pp. 1-4. IEEE, 2017.
8. T. Noi, Phan, and M. Kappas. "Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery." *Sensors* 18, no. 1 (2018): 18.
9. D. Yang, A. Alsadoon, P.W.C. Prasad, A.K. Singh, and A. Elchouemi. "An emotion recognition model based on facial recognition in virtual learning environment." *Procedia Computer Science* 125 (2018): 2-10.
10. Lyons, J. Michael, S. Akamatsu, M. Kamachi, J. Gyoba, and J. Budynek. "The Japanese female facial expression (JAFFE) database." In *Proceedings of third international conference on automatic face and gesture recognition*, pp. 14-16. 1998.
11. Gupta, Samta, and S.G. Mazumdar. "Sobel edge detection algorithm." *International journal of computer science and management Research* 2, no. 2 (2013): 1578-1583.
12. M. Chao, C. Sun, D. Song, X. Li, and H. Xu. "A deep learning approach for online learning emotion recognition." In *2018 13th International Conference on Computer Science & Education (ICCSE)*, pp. 1-5. IEEE, 2018.
13. M. Songyan, and T. Du. "Improved adaboost face detection." In *2010 International Conference on Measuring Technology and Mechatronics Automation*, vol. 2, pp. 434-437. IEEE, 2010.
14. S. Paolo, and A. Bononi. "A time-domain extended Gaussian noise model." *Journal of Lightwave Technology* 33, no. 7 (2015): 1459-1472.
15. L. Cen, F. Wu, Z.L. Yu, and F. Hu. "A real-time speech emotion recognition system and its application in online learning." In *Emotions, technology, design, and learning*, pp. 27-46. Academic Press, 2016.

16. Y. Ma, and A. Nishihara. “Efficient voice activity detection algorithm using long-term spectral flatness measure.” *EURASIP Journal on Audio, Speech, and Music Processing* 2013, no. 1 (2013): 87.
17. R. Walia, and S. Ghosh. “Design of active noise control system using hybrid functional link artificial neural network and finite impulse response filters.” *Neural Computing and Applications* (2018): 1-10.
18. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, and S. Zafeiriou. “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network.” In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5200-5204. IEEE, 2016.
19. F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions.” In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1-8. IEEE, 2013.
20. S. Mirsamadi, E. Barsoum, and C. Zhang. “Automatic speech emotion recognition using recurrent neural networks with local attention.” In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2227-2231. IEEE, 2017.
21. C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database." *Language resources and evaluation* 42, no. 4 (2008): 335.
22. A. Mojsilovic, and B. Rogowitz. “Capturing image semantics with low-level descriptors.” In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, vol. 1, pp. 18-21. IEEE, 2001.
23. H. Saif, Y. He, M. Fernandez, and H. Alani. “Contextual semantics for sentiment analysis of Twitter.” *Information Processing & Management* 52, no. 1 (2016): 5-19.
24. B. Le, and H. Nguyen. “Twitter sentiment analysis using machine learning techniques.” In *Advanced Computational Methods for Knowledge Engineering*, pp. 279-289. Springer, Cham, 2015.
25. O. Appel, F. Chiclana, J. Carter, and H. Fujita. “A hybrid approach to sentiment analysis.” In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pp. 4950-4957. IEEE, 2016.
26. Y. Li, J. Tao, B. Schuller, S. Shan, D. Jiang, and J. Jia. “MEC 2016: the multimodal emotion recognition challenge of CCPR 2016.” In *Chinese Conference on Pattern Recognition*, pp. 667-678. Springer, Singapore, 2016.
27. S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. “Convolutional MKL based multimodal emotion recognition and sentiment analysis.” In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 439-448. IEEE, 2016.
28. M. Gönen, and E. Alpaydın. “Multiple kernel learning algorithms.” *Journal of machine learning research* 12, no. Jul (2011): 2211-2268.
29. Q.T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, and A. Rehman. “Sentiment analysis using deep learning techniques: a review.” *Int J Adv Compute Sci Appl* 8, no. 6 (2017): 424.
30. P.L. Carrier, A. Courville, I.J. Goodfellow, M. Mirza, and Y. Bengio. “FER-2013 face database.” *Technical report* (2013).
31. K.V.K Kishore, and P.K. Satish. “Emotion recognition in speech using MFCC and wavelet features.” In *2013 3rd IEEE International Advance Computing Conference (IACC)*, pp. 842-847. IEEE, 2013.
32. A.T. Lopes, E.D. Aguiar, A.F. De Souza, and T. Oliveira-Santos. “Facial expression recognition with convolutional neural networks: coping with few data and the training sample order.” *Pattern Recognition* 61 (2017): 610-628.
33. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean and M. Devin et al. “TensorFlow: A system for large-scale machine learning.” In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265-283. 2016.

AUTHORS PROFILE

Dr. Archana Sharma received Doctorate from Jodhpur National University, Jodhpur, MCA from University of Jammu, India. She is currently working as Assistant Professor, Department of Computer Science, Govt.M.A.M College, Cluster University of Jammu. She has 12 years teaching experience at university of Jammu. Her research area is data mining and artificial intelligence. She has published several papers in National & International Journals.

Dr.Vibhakar Mansotra received Doctorate in computer science from University of Jammu, M.Sc., M.Phil.,(Physics), PGDCA, M.Tech. (IIT-Delhi), India. He is currently working as Professor, Former Head of Department of Computer Science and IT, Dean, Faculty of Mathematical Science & Director, CITES&M, Coordinator IGNOU(S.C-1201), University of Jammu and Chairperson Division-IV, Computer Society of India. He has 26 years teaching experience at university of Jammu. His research area is data mining, software engineering and information retrieval. He has published several papers in National & International Journals.