ISSN No:-2456-2165

Multi-class Sentiment Analysis on Multimedia Streams

[1]Piyush Singh Pasi, [2]Runal Pimpale, [3]Satyam Tiwari, [4]Bhavesh Panchal
Computer Engineering
Rajiv Gandhi Institute of Technology Mumbai, India

Abstract:- Sentiment analysis has evolved over the past few decades, Sentiment analysis has evolved over the past few decades. However, most of the work in it revolved around textual sentiment analysis. On the other hand, audio-visual sentiment analysis is still in a nascent stage. It is hard to identify an emotion based on only facial expression as it can cause ambiguity since expressions of a human face may not convey the exact emotion as their speech does. Identifying sentiments is mostly hindered by 2 classes (positive, negative) or in some cases 3 classes (positive, negative, neutral). In this proposed system, a multi-class sentiment analysis of audio and video is performed. The video comprises of 7 sentiment classes (anger, disgust, fear, happy, neutral, sadness, surprise). The audio comprises of 5 classes(anger, happy, neutral, sadness, surprise). Also, video sentiment analysis is synchronized with audio sentiment analysis to get better results. Any audio and video format supported by FFmpeg is also supported by the proposed model. The model can also be used to extract sentiments from textual data in 7 classes same as audio sentiment analysis.

Keywords:- *Natural Language Processing, Sentiment Analysis, Neural Network, Image Processing.*

I. INTRODUCTION

Sentiment analysis is the process to computationally identify and categorize opinions expressed by a particular text, speech, etc. Sentiment analysis is the study of people's emotions or attitudes towards an event, conversation on topics or in general. Sentiment analysis is used in various applications and one of them is to comprehend the mindset of humans based on their conversations with each other. For a machine to understand the mindset of the humans through a conversation, it needs to know who is interacting in the conversation and what is spoken, so sentiment analysis is performed on the data.

Understanding the mood of humans can be very useful in many instances. for example, computers that possess the ability to perceive and respond to human non-lexical communication such as emotions.In such a case, after detecting human emo- tions, the machine could customize the settings according to his/her needs and preferences. There has been researched on transforming audio materials such as songs, debates, news, political arguments to text.

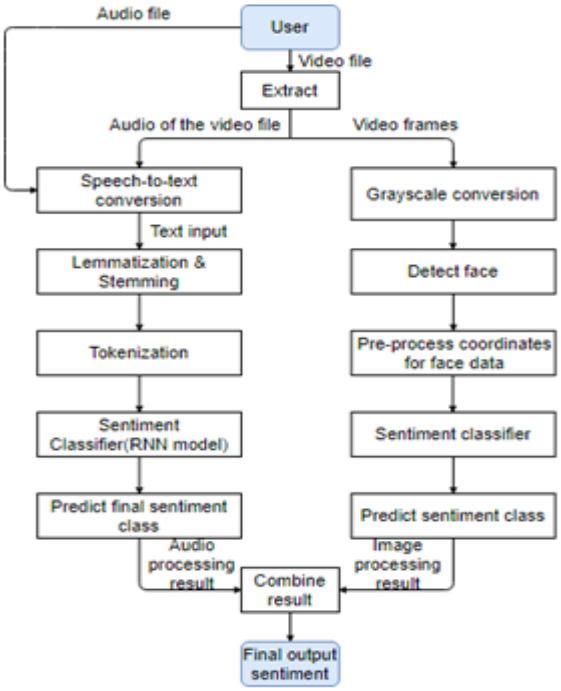

Fig 1:- The proposed system

Sentiment analysis is evolving day by day and most of the work in it revolved around textual sentiment analysis with text mining techniques. The major sentiment analysis is implemented and the sentiments are mostly limited to two or three classes.

This project aims to perform sentiment analysis on audio and video data streams by performing various intermediate process such as Speech recognition, lemmatization, tokeniza- tion, regular expression, neural network classification and using convolutional neural network model for stored video streams and in the final output sentiments can be classified into seven classes they are happy, sad, angry, disgust, fear, surprise and neutral.

II. LITERATURE REVIEW

- *Video Sentiment analysis:* [1] uses Mini Xception model which is derived from Xception [5] model. We use a similar mini Xception model as [1], however the dataset "Fer2013" used for training the Mini [1] xception model provided an accuracy of 66-percent on 7 classes "angry", "dis- gust", "fear", "happy", "sad", "surprise", "neutral". However we cleaned the dataset and reduced it to 5 classes "angry", "sad", "happy", "surprise" and "neutral" and increased the accuracy to 73-percent.

- *Audio Sentiment analysis:* [4] performs speaker specific sentiment analysis, however the classes in this proposed system is limited to 3 which are "positive", "negative", "neu- tral". However we increase the no. of classes to 7 which are "happy", "sad", "disgust", "angry", "surprise", "neutral" and "disgust", also we improve the human sentiment analysis capability of the system by implementing video sentiment analysis as well.

III. METHODOLOGY

Our audio sentiment analysis system has 7 classes of sentiments. They are happy, sad, disgust, angry, fear, neutral and surprise. And our video sentiment analysis consists of 5 sentiments, they are happy, sad, angry, neutral and surprise.

A. Audio Sentiment Analysis

Initially, the dataset is created manually this is done by collecting data from various sources, web scrapping is also used to collect data and create a dataset.

Next, pre-processing of the data is performed, this pre-processing involves lemmatization, stemming and tokenization in respective order.

Lemmatization generally refers back to the morphological evaluation of words, which ambitions to cast off inflectional endings. It helps in returning the base or dictionary form of a phrase, which is referred to as the lemma. Initially, the dictionary of words at the side of its base form is created and the given phrase matches with the phrase in the dictionary then it is replaced with its root form.

If the phrase does not exist in the dictionary then stemming rules are applied to find the root phrase, now if the word generated after stemming is valid then this phrase is considered as the new root phrase else the original phrase is retained.

The next part of the pre-processing process is tokenization. Here, we perform character level tokenization, in this proposed system the limit of the sentence size is set to 500. So, in case the size of the sentence is less than 500, the sentence is padded with zero's to reach the 500 size limit this is done so that the length of all the sentences is the same which is beneficial in the training process.

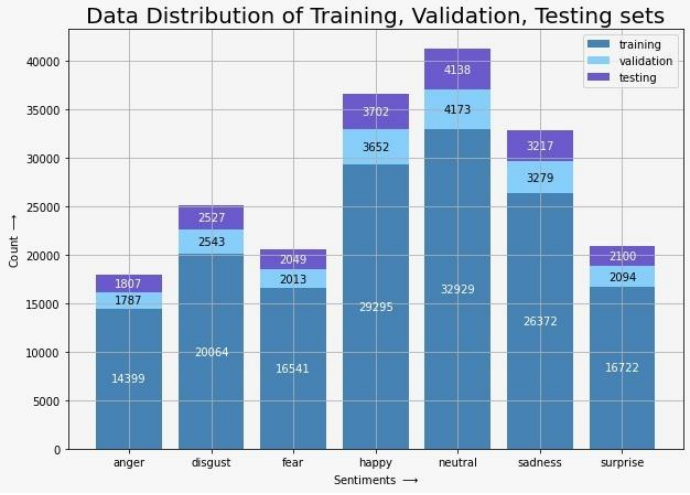

Fig 2:- Audio Data Distribution of Training, Validation and Testing set

The data is split into testing set(80-percent), training set(10- percent) and validation set(10-percent). The graph shown in Fig. 2. shows the distribution of data into training, testing and validation sets.

Now, after the pre-processing is completed the next task is to train the Neural network classifier.

While training first the embedding layer is created. In this model character level embedding is used, this is better because it can solve the out-of-vocabulary problem that is present in word-level embedding. The embedding layer is consists of a 256 output dimensions. This means that every character has 256 vector dimensions which the embedding layer selects at random. The main job of this layer is to determine the similarity between various character dimensions.

The next layer created is the conv1d layer, the kernel size that is used in this layer is 4, which means 4 rows of training data input are convoluted together, and there are 100 such filters.

Now, max-pooling is performed on the output generated by the convolution to extract maximum value. This is better understood with the figure below

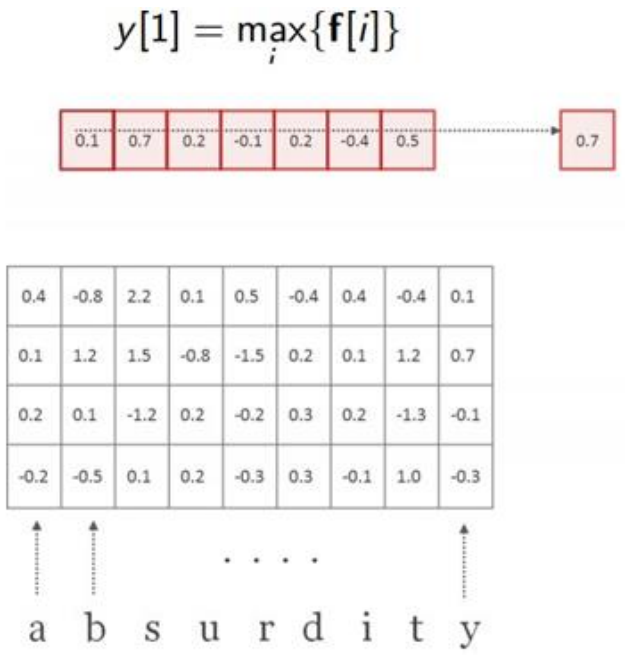

Fig 3:- Max-Pooling. Image refered from: towardsdatascience.com

Now the next layer is the bi-directional GRU layer. Bidi- rectional GRU's are a type of bidirectional recurrent neural networks with only the input and forget gates. It allows for the use of information from both previous time steps and later time steps to make predictions about the current state. The bi-directional layer consists of total 120 neurons, 60 neurons for forward direction and 60 neurons for backward direction.

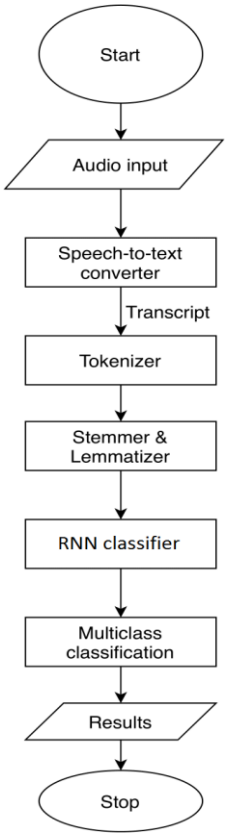

Fig 4:- Audio Sentiment Analysis Flowchart

The next layer is the dense layer, this layer consists of 64 neurons and the main task of this layer is to determine the relationships among the outputs provided by the bi-directional GRU. The activation function used in this layer is "relu"

Now, a 20-percent dropout is performed on the dense layer output to avoid over-fitting.

And finally the data is passed through the final dense layer with 7 neurons that perform the final classification, the activation function used in this layer is "softmax".

We have trained this model for 91 epochs. The loss function used is "Binary Cross Entropy", the matric used is "Categor- ical Accuracy" and the optimizer used is "Adam" optimizer.

Now, during execution when the input is provided to the first step is to convert the audio input to text form. After this pre-processing of the words is carried out on the transcript that is obtained from the input audio. The pre-processing involves lemmatization, stemming and finally tokenization.

After pre-processing is completed the data is passed through the trained model and we finally receive the sentiment classes of the input audio stream.

B. Video Sentiment analysis

In this system, we use [1] Mini Xception model to perform video sentiment analysis. The problem with this model is the dataset used as it provides only 66-percent accuracy. However, for this system, the dataset is cleaned and only the necessary emotions are retained, which includes happy, sad, angry, surprised and neutral. Which results in improvement of accuracy to 73-percent.

ISSN No:-2456-2165

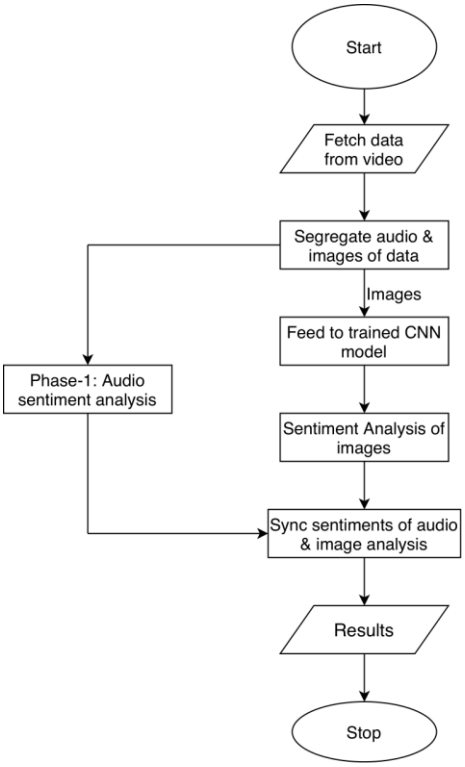

Fig 5:- Video Sentiment Analysis Flowchart

The first step is to fetch the data from stored videos. Then segregation of audio and images on the data is executed. For the audio part, the previous model is implemented and for the images, CNN sequential model is built.

Convolution neural network (CNN) is used because this net- work helps to learn multiple layers of feature representation. Preprocessing of the image is done using ImageDataGenerator and batch Normalization is also implemented on the layers. L2 regularization is used to minimize the error for loss function. Scikit-learn library is used for training and testing the models.

For input we have used “relu” activation function,and the output is generated using “softmax” optimization function. Also “Adam” optimizers are used improvised learning of the model.

The model consists of 4 hidden layers. The first hidden layer consists of 16 neurons, the second hidden layer consists of 32 neurons, third hidden layer consist of 64 neurons and finally the fourth hidden layer consists of 128 neurons, Sentiment analysis is performed in this trained model, which gives us the final sentiment class.

During execution first the video frames are extracted from video file, then these video frames are converted into grayscale images. Next, we detect the face in the video frames with the help of “haar” cascade. After this, we process the video frames so that they match with images used for training data. Now, we determine the probability of each sentiment in the frame and the highest probability determines the sentiment.

And finally the sentiment classes obtained from both audio and video sentiment analysis are used to determine the final sentiment of the input file.

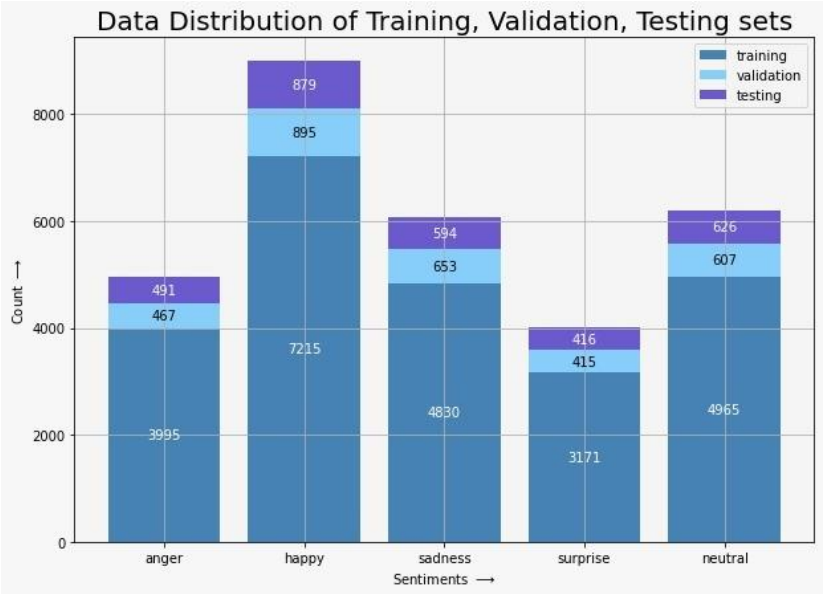

Fig 6:- Video Data Distribution of Training, Validation and Testing sets

The above given diagram represents the distribution of data among training, testing and validation sets.

The model achieved an accuracy of 73-percent with the help of 24176 (Training) + 3037 (Validation) + 3006 (Testing) = 30219 images.

IV. ABBREVIATIONS AND ACRONYMS

1)CNN stands for Convoluted Neural Network 2)RNN stands for Recurrent Neural Network 3)GRU stands for Gated Recurrent Units

ISSN No:-2456-2165

V. RESULTS

- *Audio Sentiment Analysis:* The audio sentiment analysis provides an accuracy of 75 percent with 7 sentiment classes, which are anger, disgust, fear, happy, sad, surprise and neutral.

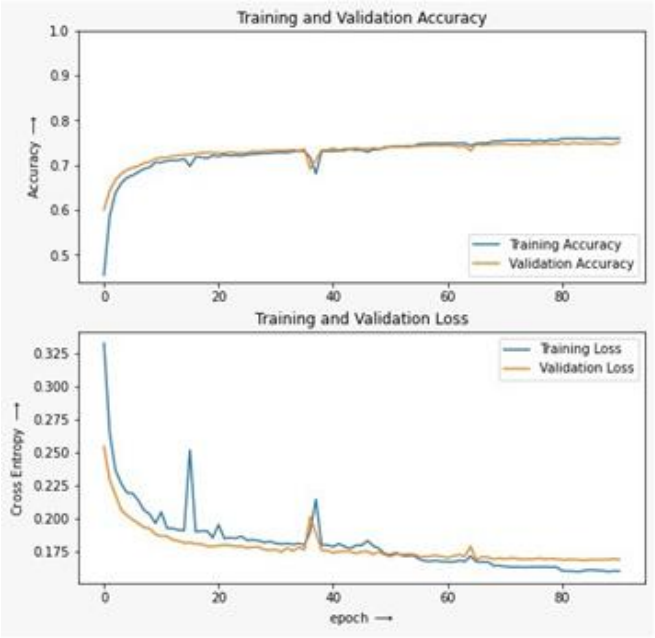

Fig 7:- Audio Sentiment Analysis model- Training and Validation Accuracy Loss curve

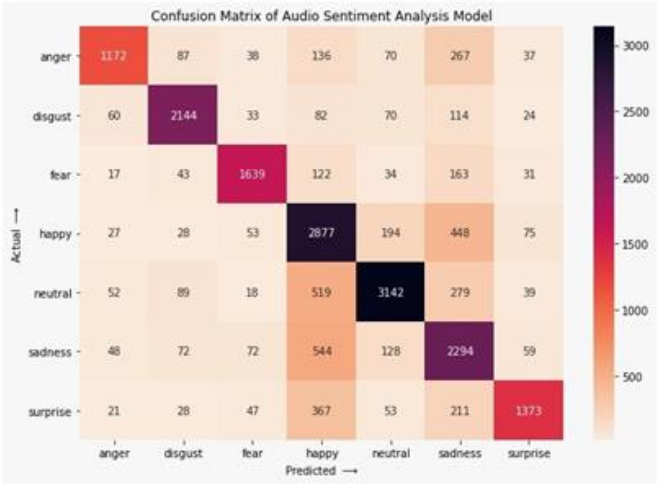

Fig 8:- Confusion Matrix for Audio Sentiment Analysis model

- *Video Sentiment Analysis:* The video sentiment analysis provides an accuracy of 73-percent with 5 sentiment classes, which are happy, sad, angry, surprise and neutral.

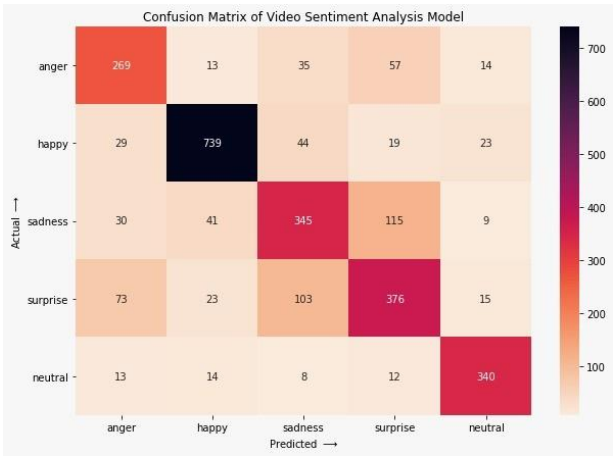

Fig 9:- Confusion Matrix for Video Sentiment Analysis model

VI. CONCLUSION

The proposed system is a generalized approach for extract- ing sentiments from multimedia streams. This method can be used for any purpose that requires multiclass sentiment analysis. The model architecture is flexible enough to be sculpted into a model that can be utilized for another specific purpose like review analysis, or classifying genres of songs, etc with minimal changes in the workflow of the system. This generalized neural network approach for multiclass sentiment analysis can be extensively used for a broad range of applica- tions.

REFERENCES

[1]. O. Arriaga, M. Valdenegro-Toro, and P. Pl¨oger, “Realtime convolutional neural networks for emotion and gender classification,” CoRR, vol. abs/1710.07557, 2017. [Online]. Available: http://arxiv.org/abs/1710.07557

[2]. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” pp. 770–778, 2016.

[3]. https://towardsdatascience.com/the-definitive-guide-to-bidaf-part-2- word-embedding-character-embedding-and-contextual-c151fc4f05bb

[4]. S Maghilnan, M Rajesh Kumar, ”Sentiment analysis on speaker specific speech data”,2017 International Conference on Intelligent Computing and Control (I2C2)

[5]. F. Chollet, “Xception: Deep learning with depthwise separable con- volutions,” CoRR, vol. abs/1610.02357, 2016. [Online]. Available: http://arxiv.org/abs/1610.02357