ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Sentiment Analysis Using Audio and Video Data

[1]Usha Mittal, [2]Pooja Rana, [3]Sanjay Kumar Singh, [4]Aditya Khamparia
[1]usha.20339@lpu.co.in, [2]pooja.20992@lpu.co.in, [3]sanjay.15745@lpu.co.in,
[4]aditya.17862@lpu.co.in
Department of Computer Science and Engineering, Lovely Professional University, Phagwara, India

Abstract
For semantic analysis, various natural language processing processes are used as sub tasks and main aim to find positive and negative sentiment in text, audio or video data. To understand the opinion of person in effective manner, vocal modulations and facial expressions can be merged to extract the features. Thus, by using multimodal methods i.e. combining text, audio and video, the performance of the system can be enhanced. In this paper, a method is proposed which uses audio and visual data to analyse semantics.

Keywords: sentiment, audio, video, convolutional neural network, Haar

1. Introduction

Notion Examination is a Characteristic Language Handling and Data Extraction task that plans to acquire speaker's emotions communicated in positive or negative remarks, questions and demands, by breaking down an enormous number of video and sound documents. As a rule, supposition investigation intends to decide the demeanour of a speaker or an essayist concerning some theme or the general tonality of an archive. As of late, the exponential increment in the Web utilization and trade of general conclusion is the main thrust behind Slant Examination today. The Internet is a colossal store of organized and unstructured information. The examination of this information to separate inert general assessment and estimation is a difficult undertaking.

The investigation of opinions might be video and sound based where the supposition in the whole video/sound is condensed as nonpartisan, quiet, upbeat, miserable, irate, frightful, disturb and shocked. It tends to be sound/video outline based where individual sentences, bearing notions, in the video/sound are arranged. SA can be express based where the expressions in a video/sound are arranged by extremity.

Estimation Examination distinguishes the expressions in each edge that bears some opinion. The individual may talk about some target certainties or abstract feelings. It is important to recognize the two. SA finds the subject towards whom the conclusion is coordinated. A casing may contain numerous substances however it is important to discover the element towards which the assumption is coordinated. It distinguishes the extremity and level of the estimation. Opinions are delegated objective (actualities), positive (means a condition of joy, joy or fulfilment on part of the speaker) or negative (indicates a condition of distress, sadness or disillusionment on part of the speaker). The opinions can additionally be given a score dependent on their level of inspiration, pessimism or objectivity.

Our everyday life has consistently been impacted by what individuals think. Thoughts and assessments of others have constantly influenced our very own sentiments. The blast of Web 2.0 has prompted expanded action in Podcasting, Blogging, Labelling, Adding to RSS, Social Bookmarking, and Long range informal communication. Accordingly there has been an emission of enthusiasm for individuals to mine these huge assets of information for conclusions. Supposition Investigation is the computational treatment of conclusions, notions and subjectivity of content. In this report, we investigate the different difficulties and uses of Slant Examination. We will examine in subtleties different ways to deal with play out a computational treatment of assessments and conclusions. Different regulated or information

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

driven strategies to SA like Credulous Byes, Greatest Entropy, SVM, and Casted a ballot Perceptron's will be talked about and their qualities and disadvantages will be addressed. We will likewise observe another component of breaking down notions by Intellectual Brain research basically through crafted by Janyce Wiebe, where we will see approaches to distinguish subjectivity, point of view in story and understanding the talk structure. We will likewise consider some particular themes in Conclusion Examination and the contemporary works in those territories**.**

2. Objective

The objective of the project is to create software that assess the performance of a speaker based on the audio and video data recorded during his/her presentation. This will help performers to deterministically analyse and rate their performance based on a combination of metrics provided by software.

3. Literature Survey:

To design software for guiding the writers, authors at Dramatica [1] had manually examined various books and films. For document recognition, in early days neural networks were widely used [2]. Now days, for visual and audio domain, deep neural networks have achieved promising accuracy in image recognition [3], speech recognition [4], and other areas.

In early days, text based sentiment analysis was the major research area ranging from short sentences to long articles. [5, 6]. There has been comparatively little work on images, with the Sentibank visual sentiment concept dataset [8] being a prominent example.

In [7], to understand main story line, multimodal descriptions are used. The more recent M-VAD [11], MovieQA [10], and Movie Description [9] works consider large-scale movie datasets that comprise a subtitles, scripts, and Described Video Service narrations.

The venture of breaking down estimations of tweets goes under the area of "Example Classification" and "Information Mining". Both of these terms are firmly related and entwined, and they can be officially characterized as the way toward finding "valuable" designs in huge arrangement of information, either consequently(unaided) or semi-naturally (administered). The undertaking would vigorously depend on procedures of "Regular Language Processing" in extricating noteworthy examples and highlights from the enormous informational collection of tweets and on "AI" methods for precisely grouping individual unlabelled information tests (tweets) as per whichever example model best depicts them.
The highlights that can be utilized for displaying examples and characterization can be partitioned into two principle gatherings: formal language based and casual blogging based. Language based highlights are those that manage formal etymology and incorporate earlier slant extremity of individual words and expressions, and grammatical features labeling of the sentence. Earlier conclusion extremity implies that a few words and expressions have a characteristic inborn propensity for communicating specific and explicit slants all in all. For instance "brilliant" has a solid positive implication while "insidious" has a solid negative meaning. So at whatever point a word with positive undertone is utilized in a sentence, odds are that the whole sentence would express a positive slant. Grammatical features labeling, then again, is a linguistic way to deal with the issue. It intends to consequently distinguish which grammatical form every individual expression of a sentence has a place with: thing, pronoun, intensifier, descriptor, action word, addition, and so forth. Examples can be separated from dissecting the recurrence circulation of these grammatical features (ether exclusively or all things considered with some other grammatical form) in a specific class of marked tweets.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

4. Proposed Methodology

It is outstanding for having the option to distinguish faces and body parts in a picture, however can be prepared to recognize practically any article. In this project, HAAR course classifiers are used for separating face from video based information and given to CNN to remove feeling from it. The calculation has four phases:

- Haar Component Determination
- Creating Indispensable Pictures
- Adaboost Preparing
- Cascading Classifiers

4.1 Haar Algorithm

Initial step is to gather the Haar Highlights. A Haar highlight considers adjoining rectangular districts at a particular area in an identification window, summarizes the pixel powers in every locale and figures the contrast between these aggregates.

4.2 Working of CNN model to extract emotional information:

a) Provide input picture into convolution layer removed from HAAR course classifier.
b) Choose parameters, apply channels with strides, cushioning if requires. Perform convolution on the picture and apply ReLU actuation to the lattice.
c) Perform pooling to decrease dimensionality size.
d) Add the same number of convolutional layers until fulfilled and straighten the yield and feed into completely associated layer
e) Output the class utilizing initiation function (Logistic Relapse with cost works) and groups pictures

Figure 1 shows the working of CNN model

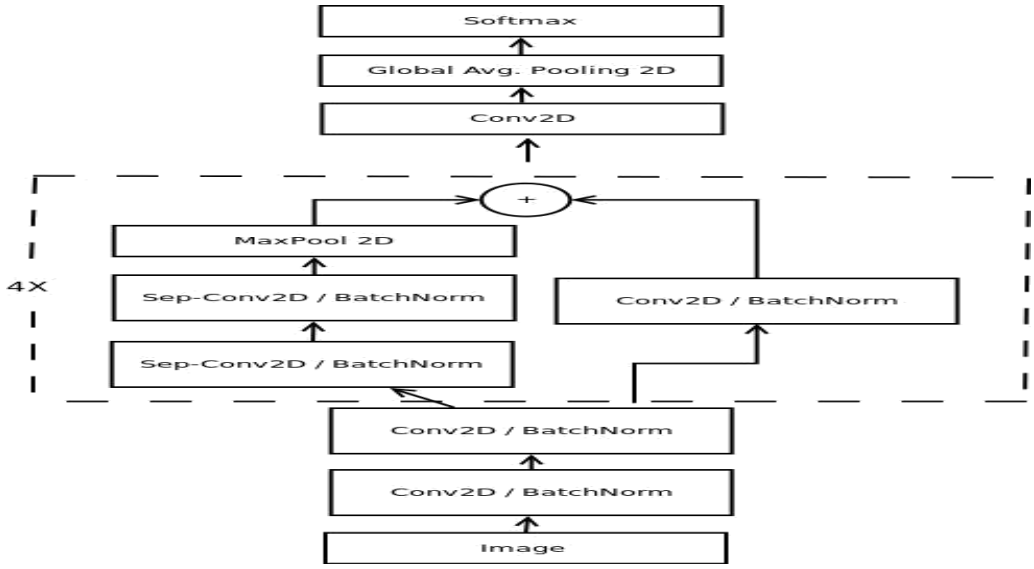

Figure 1: CNN model to extract emotional sentiment

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

4.3 Mozilla's DeepSpeech Search Algorithm:

Creation quality STT is presently the space of a bunch of organizations that have put intensely in innovative work of those advancements. To get to restrictive STT administrations, newcomers need to pay in the scope of one penny for every expression

– a cost that gets restrictive for applications that scale to a huge number of clients. To open up this zone for advancement, Mozilla plans to open source its STT motor and models so they are uninhibitedly accessible to the developer network. The Mozilla open source STT motor is intended to deal with server-class machines and can scale to serve enormous client populaces. Figure 2 shows the working of speech algorithm.

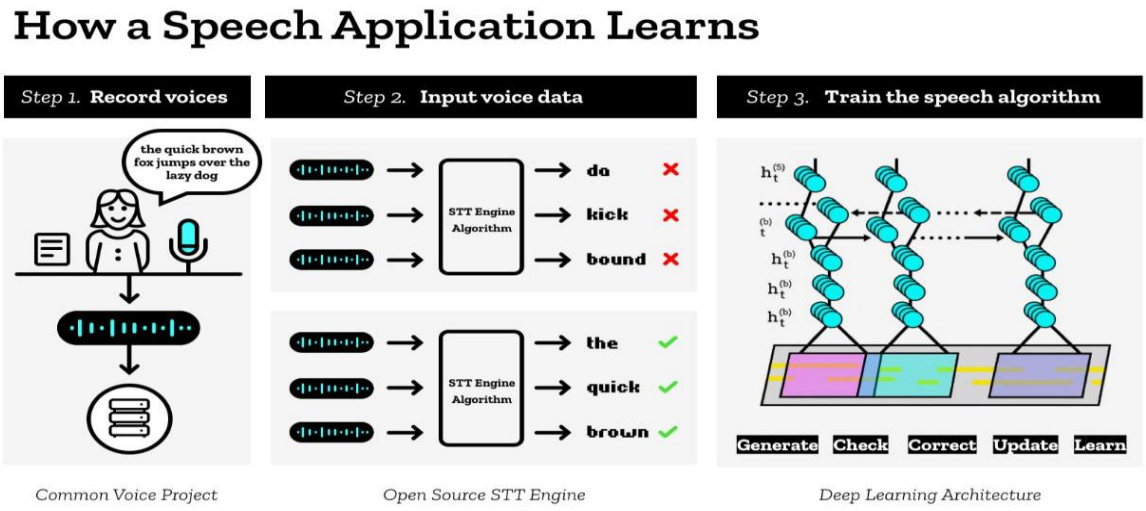

Figure 2: Working of Mozilla's DeepSpeech search Algorithm [12]

To implement the proposed system, minimum hardware requirements are GPU: with Nvidia Cuda support, Intel Core i5/i7 or equivalent and 4GB RAM. Software requirements for implementation are ubuntu/Windows, Python 3.7.4, Tensorflow Keras, cuDNN (for tensorflow-GPU), Data Visualization libraries, like matplotlib, pandas Tkinter (for GUI)
The implementation phase will be carried out in the following way:-

4.4 Push Pipe:

A custom designed python class that is used to process data in a serial manner. A PushPipe/s can be connected to another PushPipe to form a chain/tree of computation. When a PushPipe completes its processing, it calls all of its children's PushPipe.push() function with its output, thus forming a modular and easily extensible tree of computing PushPipes.

PushPipe is a base class that must be inherited to implement its PushPipe.process() function that performs the required computation. These child classes can then be used as PushPipes and connected in whatever pattern required to control data flow and perform sequential processing.

push pipe defines the following basic functions: -

PushPipe.connect(List[PushPipe]): This function accepts a list of similar PushPipe objects and appends them to the children list of the calling object. These children are sequentially

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

called by the parent (this calling object) when the parent's processing completes without any errors.

PushPipe.setErrored(errorData): This function sets the current PushPipe object to an errored state. It is used to prevent further propagation of errors when some error occurs in the processing of this object.

PushPipe.push(PassThrough): This function executes the current PushPipe's processing function and propagates its output to its children if no errors occurred. It is also responsible for profiling the processing times and catching errors as well as properly updating the PassThrough object that it receives.

PushPipe.process(data, PassThrough): This function is the one that does all the core processing for this PushPipe. It is called by PushPipe.push() and it receives two arguments, data and PassThrough. The return value of this function is stored in the PassThrough object by the PushPipe.push() function. By connecting different PushPipe to each other, the return value of one PushPipe.process() is fed to the data argument of the next PushPipe.process().

5. Results

Our project Video/Audio Sentiment Analysis is working properly. It is helping people of different audiences and people work for investigation by the results of our sentiment analysis to automate the analytical decision making process using Data Science and Machine Learning. Figure 3 to figure 8 shows some snapshots of the working system. In Figure 3, video processing is shown where emotion chart depicts the feeling of the user like happy, sad, neutral, surprise etc. Video has been taken from open source. Figure 4 shows the processing of speech i.e. audio data.

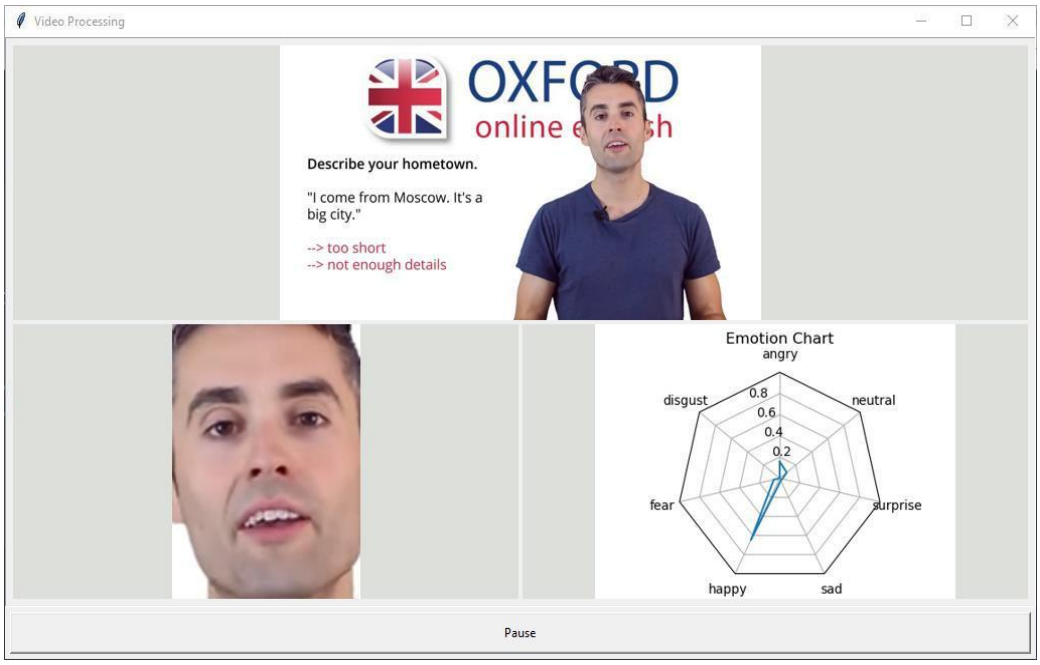

Figure 3: Video Processing

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

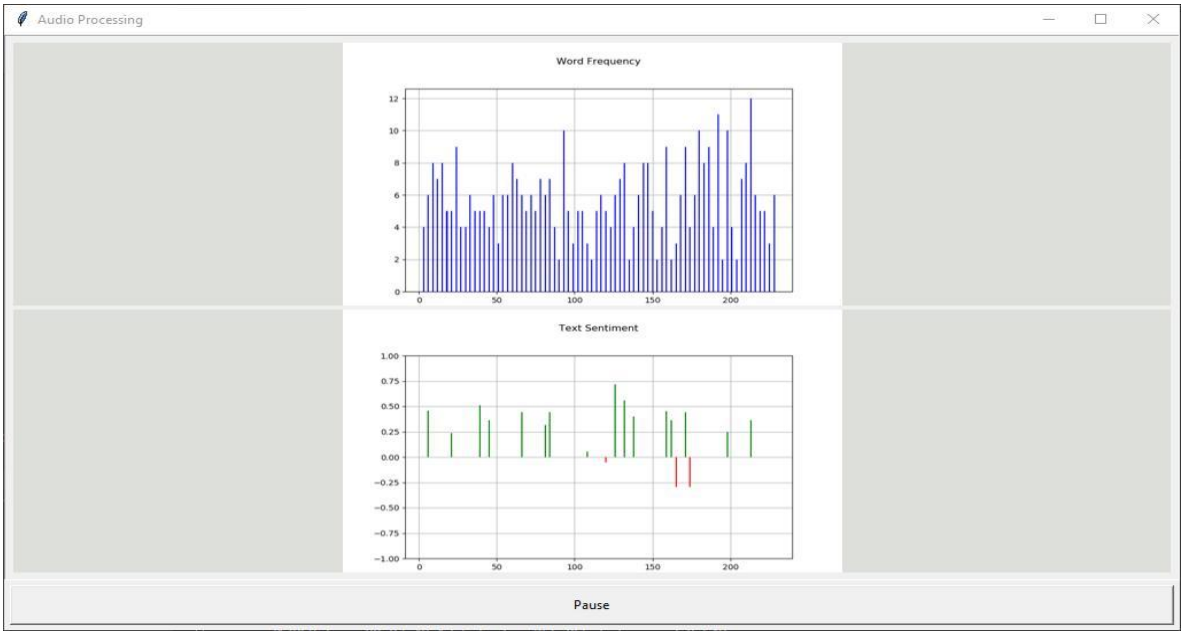

Figure 4: Audio Processing

Figure 5 shows the output video data and its statistics. Figure 6 shows the complete statistics of the video data i.e. processing time, wasted time, total time taken.

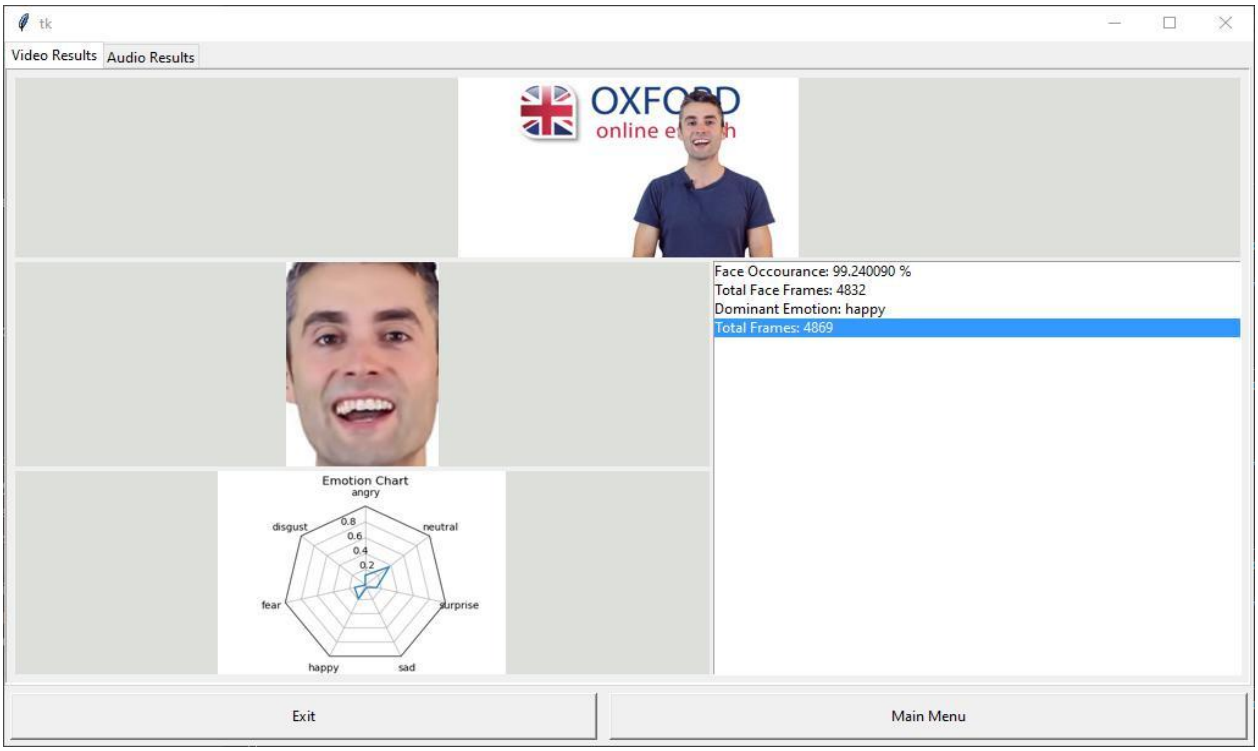

Figure 5: Output of audio and video data.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

```
Frame Statistics
Timings:
<class 'video.video_pipes.OnDemandVideoFrameExtractorPipe'> --> 0.001001 [Avg: 0.002055]
<class 'face.face_detect_pipes.FaceExtractorDNNPipe'> --> 0.075000 [Avg: 0.073775]
<class 'emotion.face_emotion_detect_pipes.EmotionExtractorOArriagaPipe'> --> 0.006004 [Avg: 0.089589]
<class 'pipedefs.core_pipes.ProcessPushPipe'> --> 0.000000 [Avg: 0.000000]
<class 'gui.gui_pipes.RadarChartRenderPipe'> --> 0.022999 [Avg: 0.027591]
Total Time: 0.105004
Process Time: 0.105004
Wasted Time: 0.000000
Efficiency: 1.000000
Worst Efficiency: 0.987711
Profiling Time: 0.010996
Worker paused.
```

Figure 6: Statistics for Video data processing

Figure 7 and 8 shows the output of audio data and its statistics.

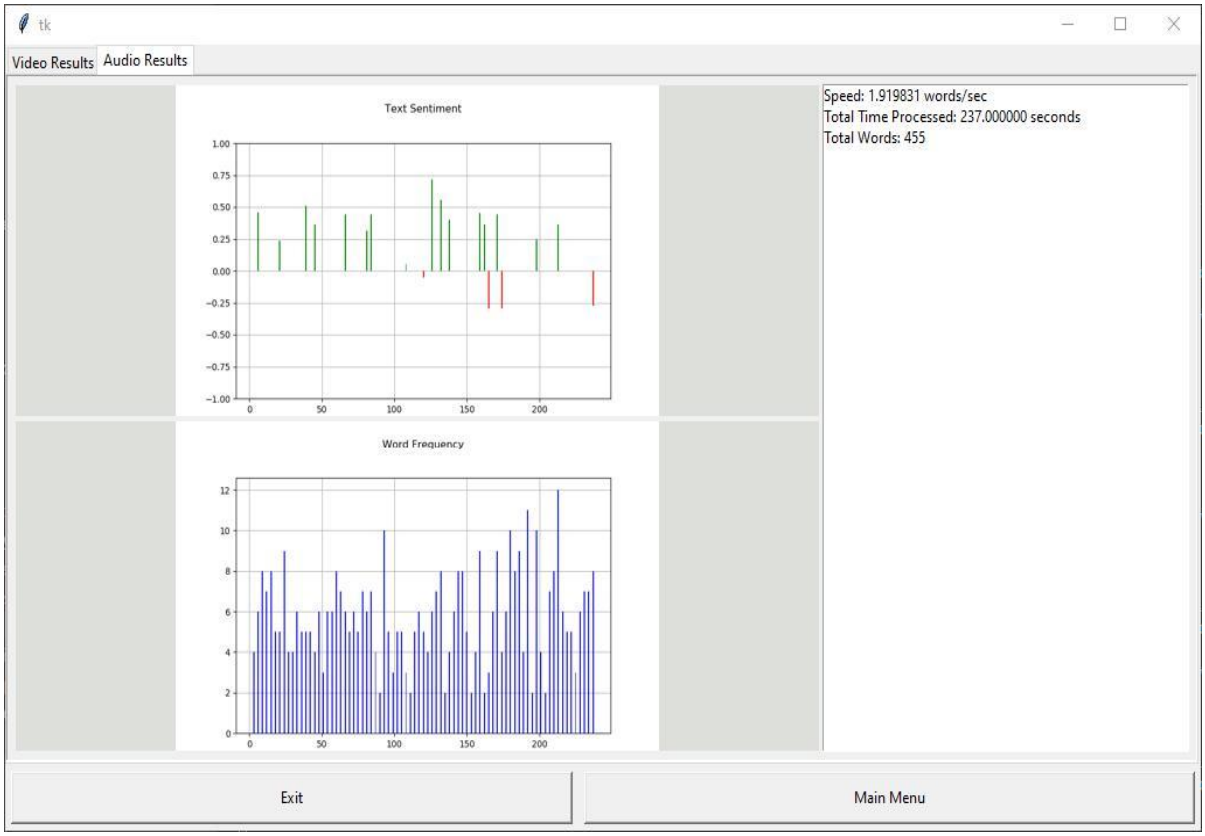

Figure 7: Output of Audio data and its statistics

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

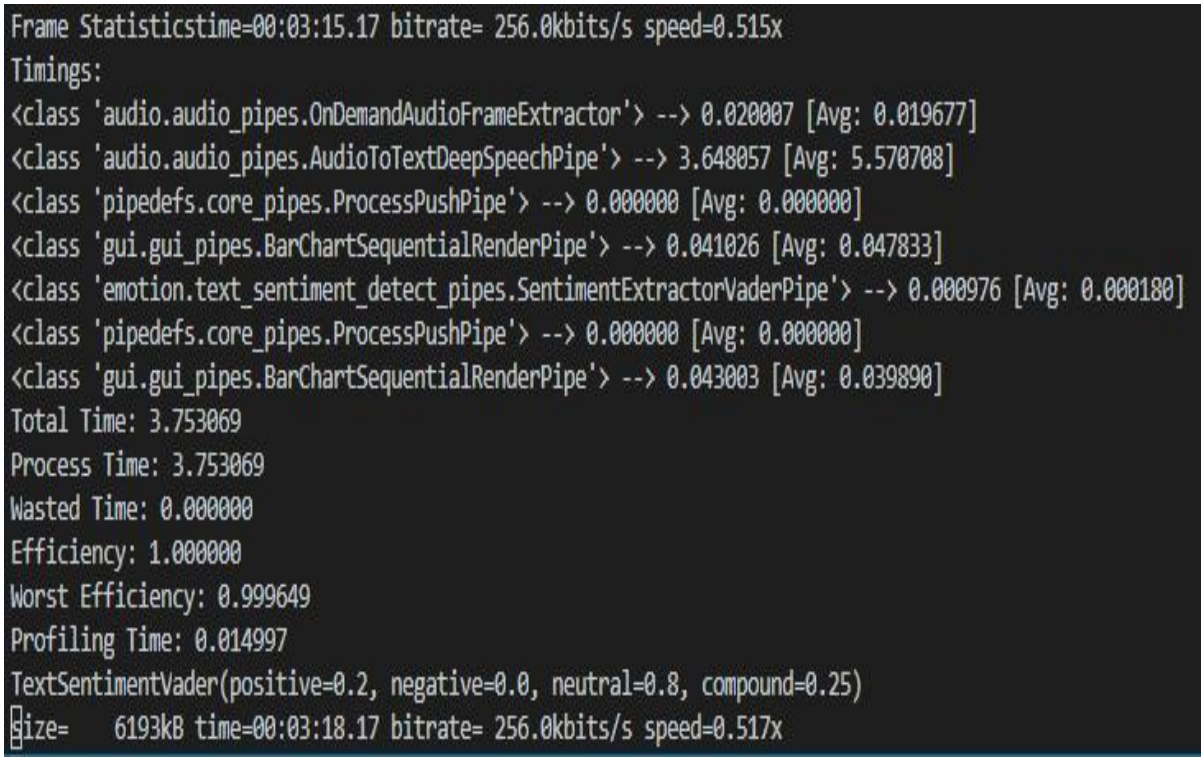

Figure 8: Statistics for Audio data processing

Results show that there is no wastage of time. Although processing time of audio data is more as compared to video data is high. But system shows very high accuracy i.e. 100% accuracy on video data and 99.9% accuracy on audio data.

6. Conclusion

As expected, bimodal model performs better than unimodal models. So far whatever is created and conveyed is only an essential successive venture which is equipped for dealing with sound and video based information sources consistently, at that point utilizing the framework it is relied upon to give exact feeling examination. Utilizing arranging framework and sending of the entire undertaking can help finding the slant examination of pre-recorded video or sound. Henceforth, it is relied upon to make increasingly cutting edge and give feeling examination progressively.

References

[1.] Dramatica: the next chapter in story development. http://dramatica.com. Accessed: 2019-11-29.
[2.] Yann LeCun, L´eon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
[3.] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
[4.] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567, 2014.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[5.] mAwni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567, 2014.
[6.] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12):2544–2558, 2010.
[7.] Andrew Salway and Mike Graham. Extracting information about emotions in films. In Proceedings of the eleventh ACM international conference on Multimedia, pages 299–302. ACM, 2003.
[8.] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM international conference on Multimedia, pages 223–232. ACM, 2013.
[9.] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description.International Journal of Computer Vision, 123(1):94–120, 2017.
[10.] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4631–4640, 2016.
[11.] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. arXiv preprint arXiv:1503.01070, 2015.
[12.] Speech and Machine Learning. https://research.mozilla.org/machine-learning. Accessed: 2019-11-19