**ABSTRACT**

The main purpose of our project Multimodal sentiment analysis is to analyze the emotions, attitude and opinions from audio visual format. A company can improve the quality of its product and services by analyzing the reviews about the product.Sentiment analysis is widely used in managing customer relations. There are many textual reviews from which we cannot extract emotions by traditional sentimental analysis techniques. Some sentences in the textual reviews may derive deep emotions but do not contain any keyword to detect those emotions, so we use audiovisual reviews in order to detect emotions from the facial expressions and by analyzing the voice of the customer .

In our project we take audiovisual input and extract emotions from video and audio in parallel from audiovisual input, finally classify the overall review as positive , negative or neutral based on combining overall emotions detected from customers facial expressions and from his voice.

**Keywords**:-Analyzing reviews, facial expression, audiovisual,

**TABLE OF CONTENTS**

**1.INTRODUCTION**

Sentiment Analysis can be many things, but in general it is a classification task. Sentiment Analysis has been one of the most targeted research topics on past decades. Given a document (e.g a review, a blog spot or a tweet), the goal is to automatically obtain its sentiment which is mostly considered as binary class problem(positive and negative) or is Multiclass problem (positive, negative and neutral).

To date, most of the works in sentiment analysis have been carried out on natural language processing. Available dataset and resources for sentiment analysis are restricted to text-based sentiment analysis only. With the advent of social media, people are now extensively using the social media platform to express their opinions. People are increasingly making use of videos(e.g. YouTube, Video, VideoLectures), images (e.g., Flickr, Picasa, Face- book) and audios (e.g., podcasts) to air their opinions on social media platforms. Thus, it is highly crucialto mine opinions and identify sentiments from the diverse modalities.

So far the field of multimodal sentiment analysis has not received much attention and no prior work has specifically addressed extraction of features and fusion of information extracted from different modalities like video and audio. Research in this field is rapidly growing and attracting the attention of both academic and industry alike. This combined with advances in signal processing and AI has led to the development of advanced intelligent systems that intend to detect and process affective information contained in multimodal sources. The majority of such state-of- the-art frameworks however, rely on processing a single modality, i.e., text, audio, or video. Further, all of these systems are known to exhibit limitations in terms of meeting robustness, accuracy, and overall performance requirements, which in turn, greatly restrict the usefulness of such systems in real-world applications.

The aim of multi-sensor data fusion is to increase the accuracy and reliability of estimates. Many applications, e.g., navigation tools, have already demonstrated the potential of data fusion. This depicts the importance and feasibility of developing a multimodal framework that could cope with all two sensing modalities: audio, and video in human-centric environments. The way humans communicate and express their emotions and sentiments can be expressed as multimodal. The audio, and visual modalities are concurrently and cognitively exploited to enable effective extraction of the affective information conveyed during communication.

With significant increase in the popularity of social media like Facebook and YouTube, many users tend to upload their opinions on products in video format. On the contrary, people wanting to buy the same product, browse through on-line reviews and make their decisions. Hence, the market is more interested in mining opinions from video data rather than text data. Video data may contain more cues to identify sentiments of the opinion holder relating to the product. Audio data within a video expresses the tone of the speaker, and visual data conveys the facial expressions, which in turn help to understand the affective state of the users. The video data can be a good source for sentiment analysis but there are major challenges that need to be overcome. For example expressiveness of opinions vary from person to person. A person mayexpress his orheropinionsmorevocallywhileothers may expressthemmorevisually.

Hence, when a person expresses his opinions with more vocal modulation the audio data may contain most of the clues for opinion mining. However, when a person is communicative through facial expressions, then most of the data required for opinion mining, would have been found in facial expressions. So, a generic model needs to be developed which can adapt itself for any user and can give a consistent result. Our multimodal sentiment classification model is trained on robust data, and the data contains the opinions of many users.

**2.LITERAL REVIEW**

Micro blogging websites have evolved to become a source of varied kind of information.This is due to nature of micro blogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain,and express positivesentiment forproducts they use in daily life.In fact,companies manufacturing such products have started to poll these micro blogs to get a sense of general sentiment for their product.Many times these companies studyuser reactions and reply to users on micro blogs. One challenge is to build technology to detect and summarize an overall sentiment.

Sentiment analysis has been a burning topic for quite a few years. Recently

it is used as a an effective tool to understand the opinions of the public and also in various social media application. With the recent growth of social websites like Facebook, Youtube and Amazon gathering public opinion i.e., honest feedback is relatively effortless.Sentiment analysis hasbeen handled as a Natural Language Processing task at many levels of granualarity. Starting from being a document level classification task,it has been handled at the sentencelevel and more recently at the phrase level.

The traditional Sentiment analysis can be done in following ways

***A. Keyword Spotting***

This approach classifies the text on basis of presence of keywords like Happy, Sad, Afraid, and Bored . This is a very naïve approach having drawback in following areas:

*a)* It can’t reliably classify negated sentences.

E.g.:- “*It was a* ***happy*** *moment” and “It wasn’t a* ***happy*** *moment”*

Both the sentences on the basis ofkeyword HAPPY will be classified as POSITIVE*.*

b) Sometimes meaning of a sentence conveys the emotion rather than any keyword in it.

E.g*.:- “My husband decided to file for divorce and hewants to take custody of my children away from me”*.

***B. Concept Based Approach***

Sentiment analysis is done on the basis of web ontology. The system grasp the conceptual and affective information associated with natural language opinions. In this approach, keywords are not used blindly but it relies on implicit meanings associated with natural language. This approach heavily relies on knowledge base it uses. Textual reviews may involve ambiguous words, for e.g.: -*bomb*, which may lead to incorrect polarities assigned to reviews.

In such cases we can have videos which contain reviews about products and on the basis of those reviews we can assign the polarity to the product. Our project sentiment analysis on audiovisual format provides an opportunity to mine opinions and sentiment. The smiles, gazes, pauses, and voice pitch are identified as relevant features. Many new areas such as facial expression, voice intensity, pauses, pitch etc. are used in opinion mining from audiovisual formats.

**4.REQUIREMENTS:**

**4.1 FUNCTIONAL REQUIREMENTS:**

***4.1.1 YouTube Dataset:***

Trusted product review video set is collected which act as major input in sentiment calculation.

**4.1 NON-FUNCTIONAL REQUIREMENTS:**

***4.2.1 Efficiency:***

As we have calculated polarities considering number of occurrences there is a chance of getting correct polarity.

***4.2.2 Accuracy:***

As far as concerned our project provides a better accuracy in regards with paper chosen.

**4.3 HARDWARE REQUIREMENTS:**

System with Windows OS and 4 GB RAM.

**4.4 SOFTWARE REQUIREMENTS:**

Matlab R2013 version

PRAAT Tool

Format Factory

**4.5 KNOWLEDGE REQUIREMENTS:**

Matlab Image processing knowledge

PRAAT tool usage manual knowledge

**5.AUDIOVISUAL FORMAT APPROACH**

***5.1 INTRODUCTION:***

Multimodal sentiment analysis is the analysis of emotions, attitude, and opinion from audiovisual format.There are many textual reviews from which we cannot extract emotions by traditional sentiment analysis techniques. Some

Sentences in the textual reviews may derive deep emotions but do not contain any keyword to detect those emotions, so we used audiovisual reviews in order to detect emotions from the facial expressions of the customer.

***5.2 TYPES OF EXPRESSIONS:***

Various features such as the location of the eyes, eyebrows and mouth are used for the analysis of facial expressions.

**T**here are three types of expressions namely:

**A. Micro expressions**

These are the expressions with the duration of 1/50 to 1/25seconds. These expressions are generated when a person feels less confidence.

**B. Macro expressions**

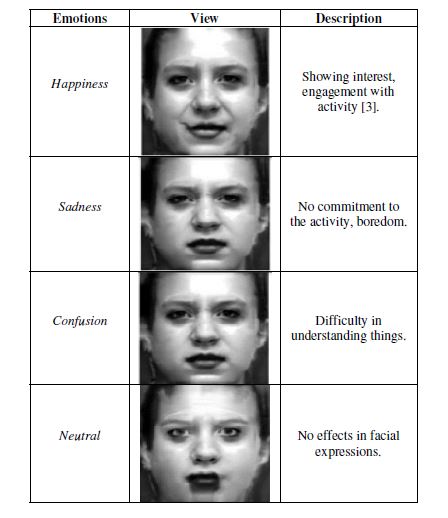
These are the expressions with the duration of 0.5 to 4seconds. These expressions are generated when a person is confident.

**C. Subtle expressions**

These expressions are generated when the intensity of emotion is not strong. In order to detect facial expressions the distance between several points like ends of eyebrows, tip of nose, ends of lips, and tip of cheeks is calculated. The change in facial expressions is detected by detecting the changes in the distance

between these points. Voice signals are used in order to increase the accuracy of emotion recognition called multimodal emotion recognition.

**Table1: Showing different Type of Expressions**



**5.3 PROPOSED WORK:**

We studied several statements that cannot be analyzed using traditional sentimental analysis techniques either due to ambiguity or multiple meaning of words. Some statements convey deep emotions but don’t contain words to

Express them.

The list of studied statements is:

1. She thought she ditched me but I am free now.

2. I met my best friend by an accident.

3. I avoided an accident.

4. I was sad because I had no shoes until I met a manwho had no feet.

5. My husband decided to file for divorce and he wantsto take custody of our children from me.

In order to overcome these ambiguities in sentiment classification we decided to switch to multimodal sentiment analysis. Multimodal sentiment analysis involve more than one signal i.e., voice and video signals in order to enhance the accuracy in classifying emotions. In our project first we have analyzed video input detected emotion and then computed the emotions from audio input and fused both the inputs in order to obtain overall polarity.

**5.4 DATA SET EMPLOYED:**

Trusted YouTube product review videos are collected and used as an input to determine the polarity sentiment. Videos in the dataset were about different topics (for instance politics, electronics product reviews, etc.).The videos were found using the following keywords: opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, baby product review, I hate, I like etc.

The videos were converted to mp4 format with a standardize of 360\_480.The length of the videos varied from 2 to 5 min. All videos were processed to avoid the issues of introductory and multiple topics. Many videos on youtube contained an introductory sequence where a title was shown sometime accompanied by visual animation. To address this issue first 30 s was removed from each video and then used for emotion detection.

**5.4 ANALYZING VIDEO INPUT:**

Preprocessed videos are received as input and then applied matlab code to convert the video frames to Image frames at 30 Hz frame rate .Emotion from individual Image frame is detected in following steps

***5.4.1 Face Detection:***

For Detecting the face in image frame we use Viola Johnes Face detection algorithm which detect the frontal face appearing in the image.

***5.4.2 Feature Points Extraction:***

After detecting the face, Using Viola Johnes Eye Detector, Mouth Detector and Nose Detector we extracted eye, lips and nose region feature points.

***5.4.3 CALUCLATING DISTANCES:***

The distance between various feature points is calculated in the baseline input image *()*.Now the distance between the points in the video image *()* is calculated and compared with the baseline image’s distance. A threshold value *(μ)* is used to classify this change in distance. Based on this difference between distances w.r.t. to threshold, the change indicator *()* is calculated as follows

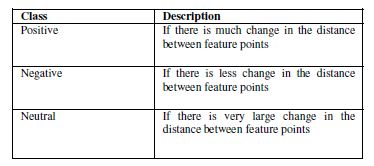
1 if -<*μ*

*= -*1 if ->*μ*

1. if - is very large

Classification can be defined as

**Table 2:Classification of Emotions**



***5.4.4 ESTIMATING PROBABILITIES:***

The probability () of classifying emotions into various classes like positive (smile, surprise),negative (sadness) and neutral (no change in emotions) depends on the value of () i.e., change estimated in feature points.

***Fig1:Change in distance between feature points w.r.t to Base Image***

***Neutral Image***

***F:\finalyear-project\jaffe\KL.NE3.157.tiff***

***F:\finalyear-project\jaffe\KL.HA2.159.tiff F:\finalyear-project\jaffe\KL.SU3.166.tiff F:\finalyear-project\jaffe\KL.DI4.173.tiff***

***Smile Surprise Sad***

Pulling at lip corners Opening of mouth shrinking of nose

and eyebrow distance

**5.5 ANALYZING AUDIO INPUT:**

To analyze audio input first the product review videos are transcribed to audio format using Format Factory software**.** Analysis of audio input involves the analysis of four features that are pitch, pause, voice intensity and loudness. These four features are analyzed using PRAAT software tool as follows

***5.5.1 PITCH:***

It measures the variation in voice during entire video. Pitch is automatically computed using PRAAT tool.

***5.5.2 INTENSITY:***

It is the measure of sound power of spoken words. Intensity is automatically computed using PRAAT tool.

***5.5.3 PAUSE:***

It is the number of audio frames counted as silent. A threshold value is used to identify the samples with or without speech.

***5.5.4 LOUDNESS :***

Strength of voice is measured as loudness. Loudness in voice is controlled by the amplitude of sound waves and measured using PRAAT tool.

**5.6 FUSION OF BOTH AUDIO AND VIDEO FEATURES**

A statement is detected by detecting a pause in the audio frames; the polarity to statement is assigned on the basis of both extracted audio and video features.

1. Positive videos are characterized by increased number of smiles and an increased number of pauses.

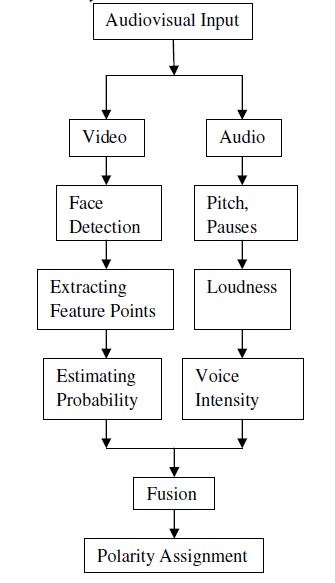
2. Negative videos are characterized by higher voice intensity and sadness over face.

3. Neutral videos are monotonous in nature and involve no change in audio and video features.

The overall polarity assigned to product is the polarityassigned to maximum number of statements in the audiovisual review i.e., positive, negative or neutral. A count is performed in order to detect polarity occurred maximum number of times.

**5.7 PROCESS FLOW CHART :**

An audiovisual review is given as input to the system and the output from the system is the polarity assigned the audiovisual review. Both audio and video signals are analyzed in parallel with the same frame rate i.e., 30 Hz with the sliding window of 50 milliseconds**.**

******

***Fig 2: Processing Flowchart of Multimodal Sentimental Analysis***