A new alley in Opinion Mining using SentiAudioVisual Algorithm

Varsha Rathi*, Mukesh Rawat**
*(Pursuing Master's in Computer Science and Engineering, Meerut Institute of Engineering and Technology, Meerut, Email:varsharathi1993@gmail.com)
** (Professor in Department of Computer Science, Meerut Institute of Engineering and Technology, Meerut Email: m_rawat1976gmail.com)

ABSTRACT

People share their views about products and services over social media, blogs, forums etc. If someone is willing to spend resources and money over these products and services will definitely learn about them from the past experiences of their peers. Opinion mining plays vital role in knowing increasing interests of a particular community, social and political events, making business strategies, marketing campaigns etc. This data is in unstructured form over internet but analyzed properly can be of great use. Sentiment analysis focuses on polarity detection of emotions like happy, sad or neutral.

In this paper we proposed an algorithm i.e. SentiAudioVisual for examining Video as well as Audio sentiments. A review in the form of video/audio may contain several opinions/emotions, this algorithm will classify the reviews with the help of Baye's Classifiers to three different classes i.e., positive, negative or neutral. The algorithm will use smiles, cries, gazes, pauses, pitch, and intensity as relevant Audio Visual features.

Keywords: Sentiment Analysis; Opinion Mining; Audio Visual Algorithm; Audio Features; Customer Reviews; Facial Expressions; Emotion Detection.

I. INTRODUCTION

Social media like Facebook, Vimeo and YouTube contain huge videos about products, services, events and interests of individuals [2]. These videos can be analyzed in order to extract opinion of individual about these products; service etc. which can help peers in decision making. If reviews are positive then peers may go for respective products and if reviews are negative then peers may opt some substitute products [13].

Social media contain a huge dataset in the form of reviews so some content level filtering technique should be used in order to filter out genuine reviews [1]. After filtering the review it is analyzed using SentiAudioVisual algorithm in order to assign polarity to it. SentiAudioVisual algorithm is implemented using programming language MATLAB[1] and PRAAT[2] application.

An audio-visual input is given to the SentiAudioVisual algorithm and this algorithm first of all separates audio and video content from each other. The audio and video input will be analyzed in parallel. Ten random frames from video are collected (the number of frames may also depends on size of video); these frames are analyzed using code written in programming language MATLAB[1]. With the help of MATLAB[1] we will extract the features like smiles, cries, etc. These extracted features will be compared with the features of Happy Set, Sad Set and Neutral Set. In order to increase the accuracy of the system we are taking differential happy postures in Happy Set, similarly differential sad postures in Sad Set, and differential neutral postures in Neutral Set. The input image is compared with each and every posture in the respective sets. The classification of input image is based on percentage matching. The input image is classified into the respective set if it has highest percentage matching to any of the posture in that set. After analyzing all ten random frames the membership values of video input for each class are obtained in Feature Vector F_1={Happy, Sad, Neutral} as follows:

$$F_1 = \begin{cases} \dfrac{No.\ of\ \ happy\ \ frames}{10}, \\ \dfrac{No.\ of\ \ sad\ \ frames}{10}, \\ \dfrac{No.\ of\ \ neutral\ \ frames}{10} \end{cases} \quad \ldots (1)$$

For example, if 6 out of 10 random frames matches maximum with Happy Set frame, 3 matches maximum with Sad Set frame and remaining 1 matches maximum with Neutral Set frame then value of Feature Vector F_1= {0.6, 0.3, 0.1}. Feature Vector F_1 describes the membership values of video input in different classes.

Audio input is analyzed using application PRAAT[2] which will identify the audio features like intensity of voice, pitch, pauses, loudness etc.[17]. These features are compared with threshold value in order to classify them as happy, sad or neutral.

Finally SentiAudioVisual merges the result obtained from both the inputs in order to assign gradual membership values to the reviews. Since SentiAudioVisual involve more than one signal and hence can assign polarity to the reviews more accurately. Final classification of the video is done by Baye's Classifiers which consider both audio, video features and results the membership values of the audio-visual input to the Positive, Negative, and Neutral classes.

II. RELATED WORK

Multimodal sentiment analysis has not been fully explored yet but has a great potential as an application. Many new areas like facial expressions, body movement, voice intensity has to be explored [3].

Facial expression features can be described with the help of thee methods namely geometry based approach, appearance based approach and combination of the two [13] . Geometry based approach classify expressions based on their deformation of facial landmark points over time. Appearance based approach uses dynamics of the texture deformation for feature extraction.

Humans communicate their emotions with the help of facial expressions [7]. Lots of work has been done in the field of face detection and face recognition, but here we will recognize the facial expressions in order to reveal prospective of a person about any product or a service [5]. In the era of Social Media, Customer always checks for past experiences of peers before spending money over similar products. This system will help in automatically detecting sentiments of peers from an audio-visual input. Active facial patches like wrinkle in upper nose region describe disgust expression and absent in other expressions [9]. Also, regions around lip corners undergo considerable changes for different expressions.

Most of the researchers detected some basic facial expressions like anger, happiness, disgust, sadness, fear, surprise etc. But in order to reduce the complexity of the system, I am dealing with only three simple expressions namely happiness, sadness and neutrality.

Not only expressions decides emotions of a person but also the paralinguistic features of speech like pitch, voice intensity are the signals can be used for affect recognition [12] . These signals used with video signals helps in accurately revealing emotions of a person. Inter-speaker variability is the main hurdle towards detecting the emotional state of a person. An emotion detection system from speech should compensate this inter-speaker variability [16]. The applications like openSMILE[3] toolkit, PRAAT[2] are used to detect speech related features like Signal Energy, Loudness, Pitch, Voice Quality etc. There should be low correlation among selected features but high correlation among selected features with emotion labels.

LBP operator is widely used as an illumination invariant feature descriptor. Neighboring pixel values are compared with the center pixel value by the operator in order to generate a binary number [22]. The pattern with 8 neighborhoods is given by

$$LBP(x,y) = \sum_{n=0}^{7} s(i_n - i_c)2^n \qquad \ldots (2)$$

Where pixel value at coordinate *(x, y)* is i_c and pixel values at coordinates in the neighborhood of *(x, y)* are i_n, and

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \quad \ldots (3)$$

The histograms of LBP image describe the feature of the image, given by:

$$H_i = \sum\nolimits_{x,y} I\{LBP(x,y) = i\}, \quad i = 0,1,....,n-1 \ \ldots (4)$$

Where *n* is the number of labels produced by LBP operator. Rotation invariant LBP operator can also be used [22].The histograms can be grouped to discover different features with the help of different bin widths. 16, 32, 256 bin histogram can be used [8].

III. PROPOSED WORK

Several videos from YouTube, Vimeo, ExpoTv, TechHive, Social sites etc. are studied for this purpose and analyzed using the SentiAudioVisual algorithm proposed in this paper. SentiAudioVisual algorithm separates audio from video and 10 random frames are chosen from the video. Then with the help of image processing like binarization, thinning etc the relevant features from each random frame are extracted, irrelevant features are filtered out and then matching is performed with the features of images in Happy, Sad and Neutral Sets [14].

Audio features are extracted using PRAAT[2], openSMILE [4] an open source toolkit. Finally features from both the signals are merged in order to classify the video into negative, positive or neutral classes.

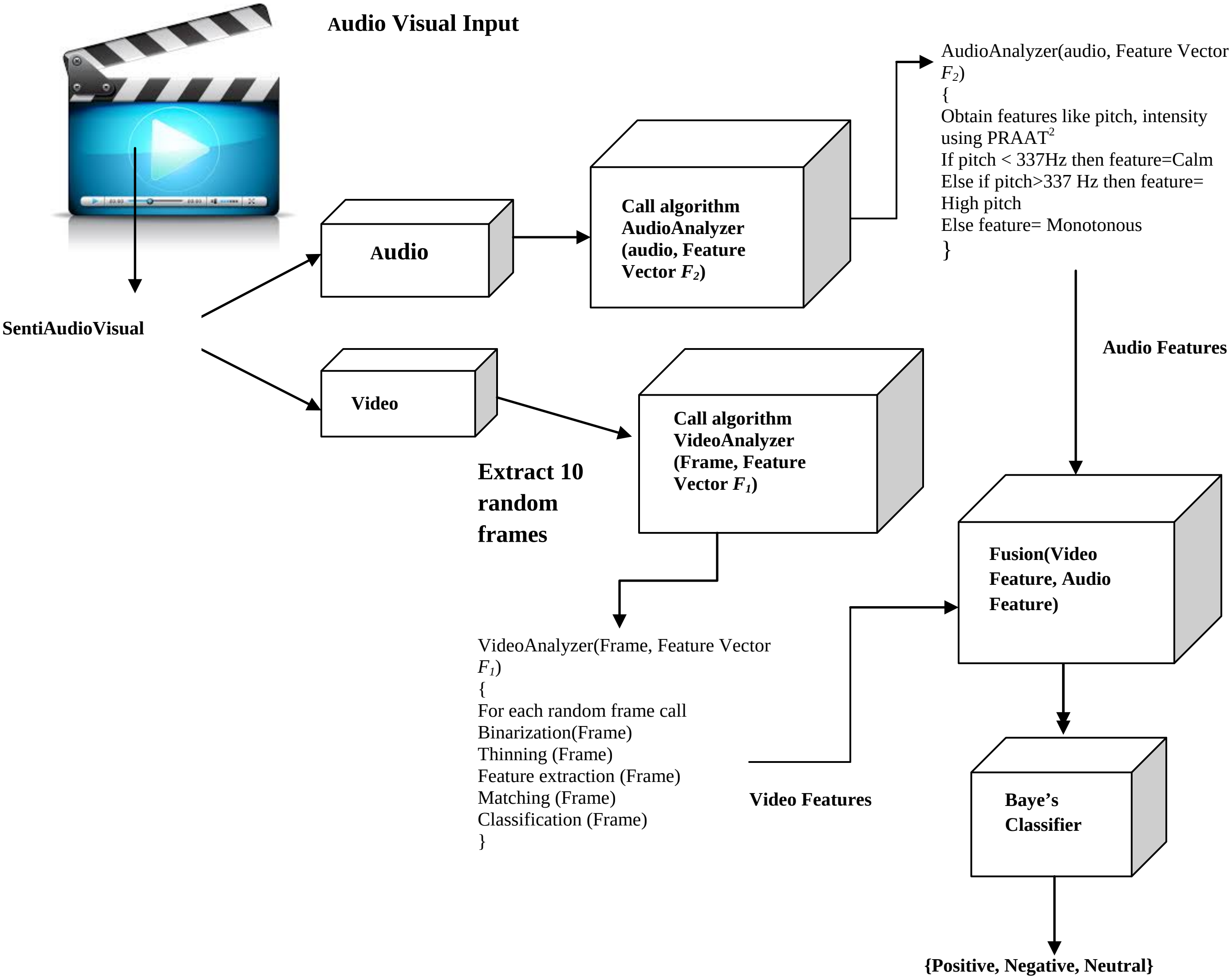

Audio Visual Input
SentiAudioVisual
Audio
Video
Call algorithm AudioAnalyzer (audio, Feature Vector F_2)
AudioAnalyzer(audio, Feature Vector F_2)
{
Obtain features like pitch, intensity using PRAAT[2]
If pitch < 337Hz then feature=Calm
Else if pitch>337 Hz then feature= High pitch
Else feature= Monotonous
}
Audio Features
Call algorithm VideoAnalyzer (Frame, Feature Vector F_1)
Extract 10 random frames
VideoAnalyzer(Frame, Feature Vector F_1)
{
For each random frame call
Binarization(Frame)
Thinning (Frame)
Feature extraction (Frame)
Matching (Frame)
Classification (Frame)
}
Video Features
Fusion(Video Feature, Audio Feature)
Baye's Classifier
{Positive, Negative, Neutral}

Fig 1. Explanation of proposed SentiAudioVisual Algorithm

MATLAB[1] Image Processing System

Some random number of frames (mostly ten) from video are obtained and analyzed to obtain facial features. Steps involved in the same are:

Binarization

It is the process of converting grey level image to binary image i.e. value 1 for active face region (facial landmarks) [21] and 0 for rest of the region.

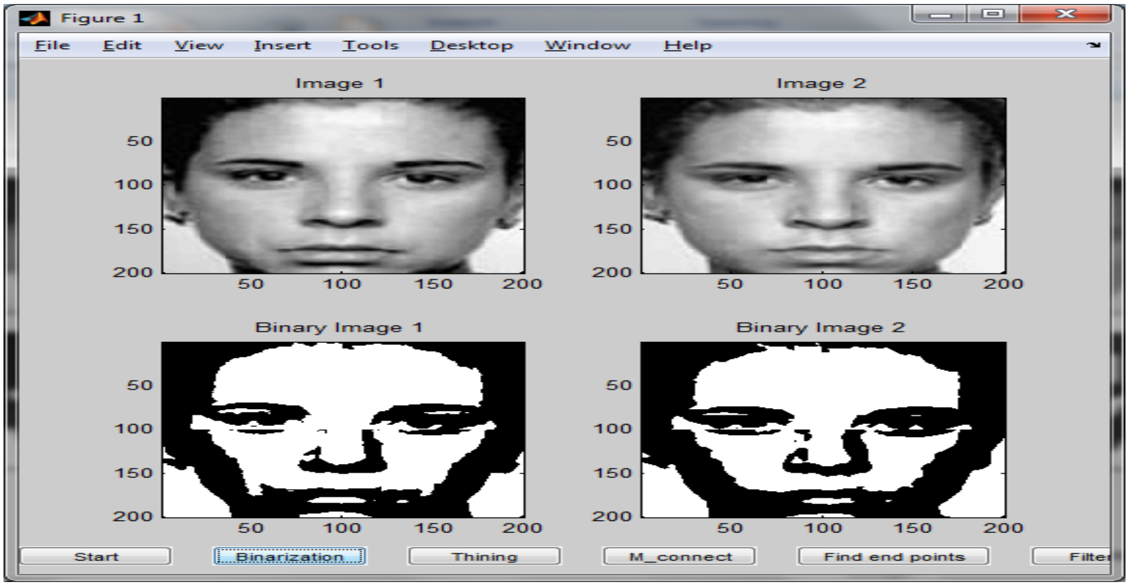


Fig 2. Binariztion performed for various posed images (Image 1) Sadness pose and (Image 2) Anger pose using (MATLAB 2010a)

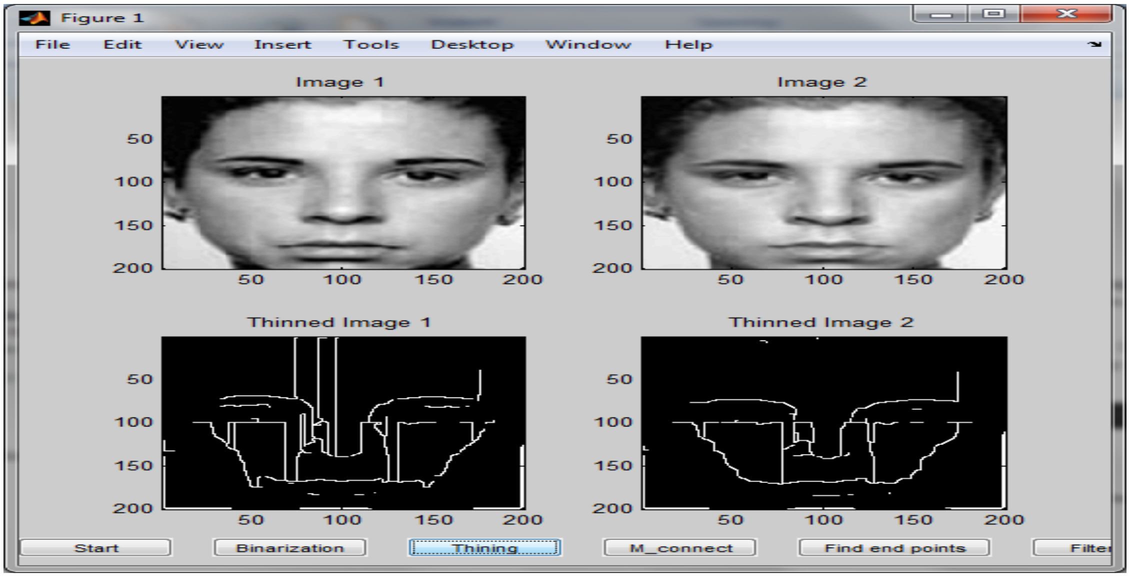


Fig 3. Thinning performed for various posed images (Image 1) Sadness pose and (Image 2) Anger pose (MATLAB 2010a)

Thinning

It is the process of converting binary image to single pixel thick skeleton. It helps in extracting feature points correctly.

Feature Extraction

Active facial patches are extracted with the help of position of eyes, eye corners, eyebrow corners, lip corners, position of nose etc .

Lip and eyebrow corners detection: Lips and eyebrow corners are detected by detecting end points in the thinned image.

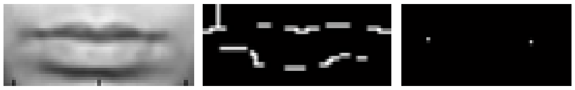
Eye corners and Nose detection: Eye corners and Nose is detected by detecting branch points in the thinned images.

Matching

While performing matching we will make three sets namely Happy Set containing at least three different happy postures, similarly Sad Set containing differential sad postures and Neutral Set containing differential sad postures. Input frame is matched against each frame of every set by comparing the distance among the feature points and percentage match is calculated [22].

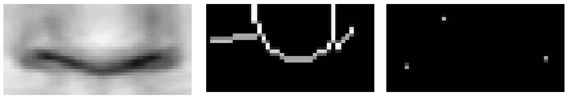
Classification

Classification is done to the set having frame with highest % match to the input frame. After classifying each frame to the respective set the membership value of visual input is calculated using Equation (1) and stored in Feature Vector F_1. Feature Vector F_1 describes the degree of membership of visual input to each class i.e. Happy, Sad, and Neutral.



(a) (b) (c)

Fig 4. (a) Lips from a face, (b) Lips after binarization and thinning, (c) Lips corner detected (MATLAB 2010a)



(a) (b) (c)

Fig 5. (a) Nose from a face, (b) Nose after binarization and thinning, (c) Nose detected (MATLAB 2010a)

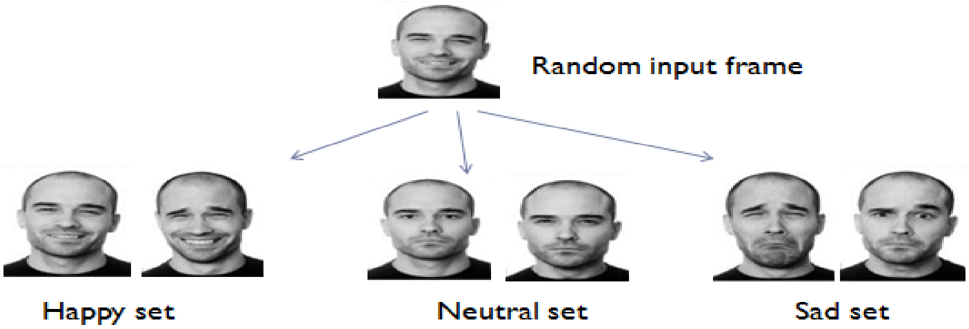


Fig 6.Matching of input frame with Different Emotions Sets (MATLAB 2010a)

Audio analysis using PRAAT[2]

Features like pitch, intensity are obtained using open source software PRAAT[2]. Value of average pitch (P_a) is obtained for entire clip. We are considering range from 75Hz to 600 Hz, which is the pitch range of normal human speech [6]. Threshold pitch is calculated using two extremes of the range.

$$Threshold \quad Pitch = \frac{75+600}{2} Hz \qquad \ldots (5)$$

Now we will compare the average pitch (P_a) of the entire clip to the threshold pitch and obtain the value of Feature Vector F_2 {Happy, Sad, Neutral}. If average pitch is less then threshold value then person is calm and happy [3]. If average pitch is more than threshold value then person is shouting and unhappy. If average pitch is equal to threshold than person is neither happy nor sad [15]. The graph in Figure 7 shows high pitch, this is the analysis of a video of product review from you tube. In this video a customer is screaming and shouting over the ill functioning of that product and suggesting peers not to go for that product. The average pitch in the video is high so value of Feature Vector F_2= {0, 1, 1} hence classified as Sad feature.

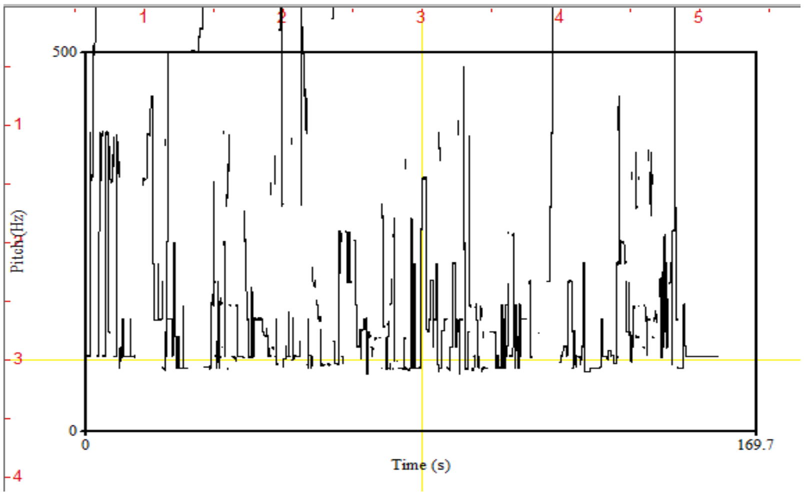


Figure 7. Audio analysis using PRAAT[2].

Fusion of both Audio and Video features and Baye's Classifier

Features obtained from both Video and Audio analyses are fused to obtain final emotion using Bayes Classifiers as follows [20] :

$$P(E_k \mid x) = \frac{1}{N}\sum_{i=1}^{N} P(E_k \mid x_n) \qquad \ldots (6)$$

Where N is the number of classifiers, x_n is the input to the classifier, $P(E_k|x_n)$ is the estimated posterior probability for class level given data, and k is the number of classes. Classifier results to the numerical value from [0 to 1] indicating the probability of given input belongs to that class. In our work we have three classes {Happy, Sad, Neutral} so values of k ranges from 1 to 3, two classifiers {Audio, Video} so value of

N= 2, and estimated posterior probabilities in the Feature Sets F_1 and F_2.

Hence membership value to Happy class can be calculated as:

$$P(E_1 \mid x) = \frac{1}{2}\sum_{i=1}^{2} P(E_1 \mid x_n) \ \ldots (7)$$

Similarly for other classes also membership value is calculated. The results of the classifier are in the range from [0 to 1]. The membership values for all the three classes are obtained [10]. Finally based on the maximum membership value the audio-visual customer review is classified as Positive, Negative, and Neutral.

SentiAudioVisual Algorithm:

Input: AudioVisual Input

Output: Sentiment class (Positive, Negative, Neutral)

Steps:

1. Separating audio and video from audio-visual input.
2. **For Video Analysis:**
   Call VideoAnalyzer(Frame, Feature Vector *F*$_1$)
   - i. Obtain 10 random frames from video.
   - ii. For each obtained frame call the following functions :
   - iii. **Binarization (Frame)**
     Threshold value is calculated using following formula:

$$[w, h] = size(Frame) \qquad \ldots (8)$$

$$sum = \sum_{i=0}^{w}\sum_{j=0}^{h} pixel(i, j) \ldots (9)$$

$$Threshold\ \ T_0 = \frac{sum}{w \times h} \ldots (10)$$

```
For each pixel in frame
    If pixel (i, j)>$T_0$
        Then pixel (i, j) =1
    Else
        pixel (i, j)=0
//convert input frame into binary form (black and white)
```

iv. **Thinning (Frame)**

```
Considering 3× 3 neighborhood for each pixel in input frame
    If pixel has only one foreground neighbor do not delete it.
    If pixel has less than equal to five foreground neighbors do not delete it.
    If pixel is a connected component do not delete it.
    Else delete the pixel.
```

v. **Feature Extraction (Frame)**

```
End points and Branch points are detected for each pixel $p_i$
is_end($p_i$)
    Sum of 8 nearest neighbor pixels is calculated.
    If (Sum==0|| Sum==1)
        Then $p_i$ is end point
    Else
        $p_i$ is not end point
is_branch($p_i$)
    Sum of 8 nearest neighbor pixels is calculated.
    If (Sum==3)
        Then $p_i$ is branch point
    Else
        $p_i$ is not branch point
```

vi. **Matching (frame)**

Comparing distances between extracted feature points of input frame to the feature points of stored frames in each of three sets i.e.; Happy Set, Sad Set and Neutral Set.

$$\%Match = \frac{No.\ \ of\ \ Matches \times 100}{Total\ \ Points} \ldots (11)$$

vii. **Classification (Frame)**

```
For each input frame:
Assigned set= Set containing frame with maximum %match.
After analyzing each frame:
Feature Vector $F_1$ = {Happy, Sad, Neutral} is calculated using Equation (1):
```

$$F_1 = \begin{cases} \dfrac{No.\ \ of\ \ happy\ \ frames}{10}, \\ \dfrac{No.\ \ of\ \ sad\ \ frames}{10}, \\ \dfrac{No.\ \ of\ \ neutral\ \ frames}{10} \end{cases}$$

F_1 contain the membership values of video input in different classes.

3. **For Audio Analysis:**
Call AudioAnalyzer(Audio, Feature Vector F_2)

```
If average pitch ($P_a$) of the entire clip is computed using PRAAT[2].
Threshold Pitch is 337Hz from Equation (5).
Feature Vector $F_2$= {Happy, Sad, Neutral} is calculated as:
    If ($P_a$ < 337Hz)
        Then $F_2$= {1, 0, 0}
    Else if ($P_a$ >337 Hz)
        Then $F_2$= {0, 1, 0}
    Else $F_2$= {0, 0, 1}
```

4. **Fusion(video features F_1, audio features F_2):**

Final classification is done using Baye's formula in Equation (6):

$$P(E_k \mid x) = \frac{1}{N}\sum_{i=1}^{N} P(E_k \mid x_n)$$

Where $P(E_k|x_1)$ = F_1 = Video Features, $P(E_k|x_2)$= F_2 = Audio Features, N=2 (Number of classifiers), k=3 (Number of classes).

For Example, if F_1 is calculated as {0.6, 0.3, 0.1} and F_2 is calculated as {1, 0, 0}. Then membership value for Happy class=

$$P(E_1 \mid x) = \frac{1}{2}\sum_{i=1}^{2} P(E_1 \mid x_n)$$

$$= \frac{0.6+1}{2} = 0.8$$

Similarly, for Sad class $P(E_2|x)$=0.15, and for Neutral class $P(E_3|x)$=0.05.

Based on maximum membership value review is finally classified as:

```
    If max (Happy)
       Class=Positive
    If max (Sad)
       Class=Negative
 If max (Neutral)
     Class=Neutral
```

IV. EXPERIMENT AND RESULT IN JUXTAPOSITION WITH OTHER TECHNIQUES:

For experimental purpose we analyzed 30 videos of customer reviews as mentioned in Table 1. These videos are collected from different social sites, shopping sites, weblogs and YouTube, Expotv etc. The results of 23 observations out of 30 were favorable, which is approximately 76.34% shown by confusion matrix in Table 2. Favorable results are those results which are accurately classified to the class they belong to. Unfavorable results are those results which are wrongly classified to the class they don't belong to.

The accuracy of proposed system is compared with the accuracy of other sentiment analysis approaches in Figure 8 which is much higher than other opinion mining techniques.

Table 1. Data Set of analyzed videos

Videos	Link
Laptop Review	https://www.youtube.com/watch?v=QF9x3ArzZ3k
Movie Review	https://www.youtube.com/watch?v=IwjFZWEmw0c
Beauty Product Review	https://www.youtube.com/watch?v=8DeMbKWHBR8
Car Review	https://www.youtube.com/watch?v=ww1YkmArX70
Water Purifier	https://www.youtube.com/watch?v=VzgG7LTodUs

Table 2. Confusion matrix for analyzed videos

			Truth	
	Classes	Happy	Sad	Neutral
Prediction	Happy	**10**	1	1
	Sad	0	**6**	1
	Neutral	2	2	**7**
Accuracy Rate for each class		0.833	0.667	0.77
Total Accuracy			**76.34**	

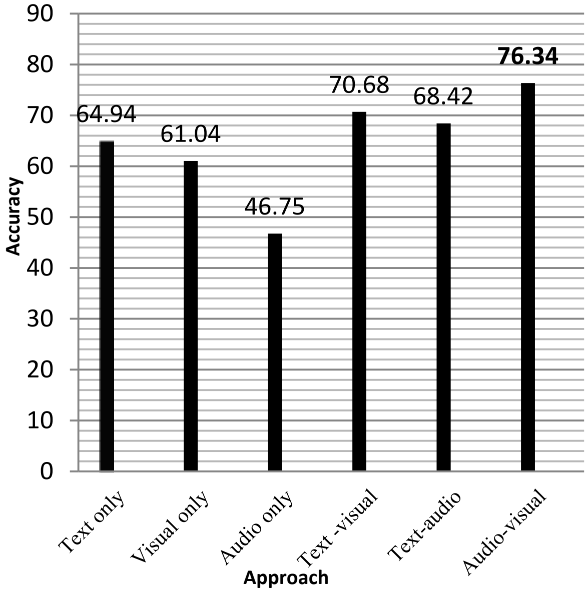


Fig 8. Comparison of Audio-visual approach with other sentiment analysis approaches

V. CONCLUSION

SentiAudioVisual algorithm automatically classifies customer's reviews as Positive, Negative, or Neutral and hence helps the peer customers to go for that product or not. The algorithm is multimodal means uses more than one signal in order to increase the accuracy of classification over the other sentiment analysis techniques. Several audio and video features like smiles, sadness, pitch, intensity are used in order to detect the emotions of a customer. Baye's classifier utilizes both audio and video features in order to declare a customer review as positive, negative or neutral. The complexity of system is reduced since it considers basic audio-video features which are easy to detect and analyze.

VI. FUTURE WORK

Future work involves customizing the proposed system according to the size of videos in which we can consider the variable number of input frames to VideoAnalyzer module. Also we can include some other audio features like pauses etc. in order to enhance the classification. The existing system can be extended to analyze both audio and video signals simultaneously unlike this system where we have separated audio-video signals and analyze them sequentially. A parallel processing of audio-visual signals will reduce the time of review classification. Hence will results into a faster sentiment analysis system.

REFERENCES

Journal Papers:

[1] A-Rong Kwon, Kyung-Soon Lee. (2013,July-Sept). *Opinion Bias Detection with Social Preference Learning in Social Data*. International Journal on Semantic Web and Information Systems IGI Global, 9(3), 57-76.

[2] Andrés García-Silva, Víctor Rodríguez-Doncel, Oscar Corcho. (2013, July-September). *Semantic Characterization of Tweets Using Topic Models: A*

Use Case in the Entertainment Domain. International Journal on Semantic Web and Information Systems, IGI Global, 9(3), (pp. 1-13).

[3] Liu, Bing. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

[4] Cambria, E., and Hussain, A. (2012). *Sentic Computing: Techniques, Tools, and Applications, Springer*.

[5] B. Schuller et al. (2011). *Recognizing Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge*. Speech Comm., vol. 53, nos. 9/10, (pp. 1062–1087).

[6] Batliner, A., Buckow, A., Niemann, H., N€oth, E., and Warnke, V. (2000). *The Prosody Module*. VERBMOBIL: Foundations of Speech-to-Speech Translations, Maybury, M., Stock, O., Wahlster, W. eds., (pp. 106-121), Springer Verlag.

Books:

[7] Liu, Bing. (2010). *Sentiment analysis and subjectivity. Handbook of natural language processing 2*. (pp. 627-666).

Proceedings Papers:

[8] Happy, S.L., Aurobinda Routray. (2015, Jan-March). *Automatic Facial Expression Recognition Using Features of Salient Facial Patches*. IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 6, NO. 1.

[9] Lago, P., & Guarín, C.J. (2014). *An Affective Inference Model based on Facial Expression Analysis*. IEEE Latin America Trans, Vol. 12, No. 3.

[10] Tayal, D.K., Sumit Yadav, Komal Gupta, Bhawna Rajput, & Kiran Kumari. (2014). Polarity detection of sarcastic political tweets. *In proceedings of International Conference on In Computing for Sustainable Global Development (INDIACom), IEEE*. (pp. 625-628).

[11] Mouthami, K., Nirmala Devi, K., Murali Bhaskaran, V., (2013). Sentiment Analysis and Classification Based On Textual Reviews. *In proceedings of International Conference on Information Communication and Embaded System (ICICES)*. (pp. 271 – 276).

[12] Carlos Busso, Angeliki Metallinou, & Shrikanth Narayanan. (2013, Oct-Dec). *Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech*. IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 4, NO. 4.

[13] Cambria, E., Schuller, B., Yunqing Xia, Havasi, C. (2013). *New Avenues in Opinion Mining and Sentiment Analysis. Intelligent System, IEEE, Vol. 28*. (pp. 15-21).

[14] Rosas, V.P., Mihalcea, R., Morency, L. (2013). *Multimodal Sentiment Analysis of Spanish Online Videos. Intelligent System, IEEE, Vol. 28*. (pp. 38-45).

[15] Songfan Yang, & Bir Bhanu. (2012). *Understanding Discrete Facial Expressions in Video Using an Emotion Avatar Image*. IEEE Trans. Systems, Man, and Cybernetics—Part B: Cybernetics, Vol. 42, No. 4.

[16] Silvia Monica FERARU. (2012, May). *Emotional Speech Classification for Romanian Language - Preliminary Results*. 11th International Conference on DEVELOPMENT AND APPLICATION SYSTEMS, Suceava, Romania.

[17] Lizhen Liu, Xinhui Nie, & Hanshi Wang. (2012). Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis. *In proceedings of 5th International Congress on Image and Signal Processing IEEE*. (pp. 1620 – 1624).

[18] Lei Zhang, & Bing Liu. (2011). Extracting Resource Terms for Sentiment Analysis. *In proceedings of the 5th International Joint Conference on Natural Language Processing*. (pp.1171-1179).

[19] B. Lu et al. (2011). Multi-Aspect Sentiment Analysis with Topic Models. *In Proceedings of Sentiment Elicitation from Natural Text for Information Retrieval and Extraction. IEEE CS*. (pp. 81–88).

[20] Chul Min Lee, & Shrikanth S. Narayanan. (2005, March). *Toward Detecting Emotions in Spoken Dialogs*. IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 13, NO. 2.

[21] Ekman, P., Friesen, W.V., & Hager, J.C. (2002, May). *FACS Manual, Salt Lake City, UT, USA: A Human Face*.

[22] T. Ojala, M. Pietikainen, and T. Maenpaa. (2002, July). *Multiresolution grayscale and rotation invariant texture classification with local binary patterns*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, (pp. 971–987).

ENDNOTES

[1] In this research, MATLAB 2010a is used.
MATLAB and Statistics Toolbox Release 2010a, The MathWorks, Inc., Natick, Massachusetts, United States.
Maltab: http://wordnet.princeton.edu/

[2] For audio analysis PRAAT is used.
PRAAT: http://www.fon.hum.uva.nl/praat/

[3] openSMILE: http://www.fon.hum.uva.nl/praat/

[4] For face Detection and Normalization FACE++ is used.
Face++: http://www.faceplusplus.com/

[5] https://www.mashape.com/apicloud/facerect