Survey of Approaches for Building Emotion Detection Applications using a Multi-modal Approach

Siddhesh More
PCCOER
Ravet, Pune

Shruti Subramanyam
PCCOER
Ravet, Pune

Kalyani Reddy
PCCOER
Ravet, Pune

Komal Patil
PCCOER
Ravet, Pune

Archana Chaugule, PhD
PCCOER
Ravet, Pune

ABSTRACT
Emotion detection of users is a challenging and exciting field where user's data is analyzed to recognize emotions such as happy, sad, angry etc. This data could be in one or multiple formats such as audio, video, text, still images etc. Relevant features are extracted and fused together to give a label. Fusing data from two or more sources(modalities) is another challenge, feature level or decision level fusion is employed. This paper inspects and studies the various approaches to multi-modal extraction of emotions.

General Terms
Emotion recognition, Sentimental analysis, Data Science, Knowledge engineering

Keywords
Sentimental analysis, Emotion detection, Multi-modal approach, Decision level fusion, Text mining, Image Classification

1. INTRODUCTION
In today's day and age smart intelligent systems relying on machine learning have become ubiquitous. Everything from search engines, recommendation system on online shopping websites to personal assistants, fitbits and of course smartphones, various day to day objects are being empowered by a machine learning algorithm that lends it the ability to make decisions based on past experiences to accomplish its goal. All this though doesn't consider the mood of the user into account. Creating computer applications that are empathetic to us, i.e understand, analyze and respond to human emotions definitely would be better at providing solutions. Therefore, we look into detecting the emotion of a user using their tweets and facial expressions and combine it with the answers to a beck decision inventory questionnaire.

Scientists at UC Berkeley have identified 27 distinct human emotions. These emotions are associated with a wealth of information about the human mind. Equipping computers to recognize emotions will have benefits in various fields. As aides to psychologists in diagnosing depression,in designing better products that connect well with customers needs,to develop smart tutoring systems[1] that teach in relevance to a student's learning ability, an autonomous car system which can recognize tired river and switch to autopilot , a personal assistant that can understand tone of user etc. A multi-modal approach that is any combination of text, audio, visual, body posture, hand gestures, facial expressions etc would give a comprehensive insight into the user's mind. A single modality may give only one sided info or may miss out on an inherent parameter. A lot of literature exists that combines audio and video data [2][3][4]. Thus a system is proposed combining text(tweets) and video features[5] to predict user's emotion as this combination gave a better result than any other[1]. Since Twitter is the most popular micro-blogging site which is regularly and frequently used to express sentiments it was our ideal choice as a source for text, video data is taken in real time as user answers the beck 9 questionnaire.

The next point of focus is extracting suitable features and combining them to predict the emotion label. There are two popular approaches to combine feature sets which have been compared 1) decision level 2) feature level[2][6][7].

The remainder of the paper describes the following sections 2. Related Literature 3. Text-based analysis 4. Video-based Analysis 5. Proposed System 6.Conclusion

2. LITERATURE SURVEY
There is vast literature that explores a multi-modal approach for detecting the emotion of a user. They all share the common theme of acknowledging the difficulty in understanding the subjectivity of human emotions and accuracy of understanding the context of interaction at the time in which the emotion was expressed. All papers seem to address the common shortcoming of a limited dataset, and that is why the multi-modal approach is the best as they outperform systems where only a single modality is chosen. Since a single modality may miss out on an inherent parameter for depression detection or may not give full information.

In a paper[1] authored by Rahul Gupta, Nikolaos Malandrakis, and Bo Xiao proposed system combined features extracted from the text, video, and audio to predict depression. All features were linearly combined and classified using a support vector regression (SVR).A multi-stage feature selection process was adopted, a brute force strategy is applied to extract a subset of feature groups. Then a best-rst forward search strategy on the combination of features obtained.

This paper also outlined the results of various combinations of features, video, audio and, text.As evident from the experiments, a combination of video and text gave the best result with feature level fusion.Ramon, Mara, Gilberto propose a tutoring system[1] that extracted the current emotion of student using video and text from the chat box. The results from both these modalities along with other parameters like time to solve question and error rate were combined with fuzzy logic to determine next level

Another work by Carlos, an audio-video bimodal system classified data into 4 categories of anger, happiness, sadness,

and surprise. Feature level fusion had an almost equivalent accuracy of 89% when compared to the decision level fusion. Both superseded the single individual modality trained classifiers in terms of performance. As evident from the confusion matrix, a feature-based bi-modal system performed slightly better than a decision-based system. The authors of another paper[3] Simina Emerich, Eugen Lupu, Anca Apatean first identified the variously suitable feature to classify data into 6 emotions namely sadness, happiness, anger, disgust, fear and neutral. It used an SVM classier employed with an RBF-kernel was used as it gave better results than other classifiers such as k-means and Naive Bayes. It followed two approaches for feature integration one was feature level fusion wherein a single feature vector was formed and normalized using z-score transformation and a match score method. The feature level fusion gave better results in correctly identifying emotions with 93% which was 1% more than match score.

Emotion Recognition System / Technique		Speech Information	Facial Expressions	Feature Level Fusion
Classifier	10-fold cross validation			
	SVM (RBF)	87.7%	90.3%	93%
	SVM (POLY)	85.2%	88.8%	90.2%
	Naïve Bayes	67%	68.14	68.7%
	K-NN (k=3)	73.7%	84.4%	86.6%
	80% training 20% testing			
	SVM (RBF)	83.3%	86.7%	91.1%
	SVM (POLY)	83.3%	85.2%	88.8%
	Naïve Bayes	66.6%	70.3%	70.7%
	K-NN (k=3)	72.2%	83.3%	85.2%

Figure 2.Results of Feature level fusion[3]

Liyanage C. De Silva, Pei Chi Ng, of Singapore used statistical techniques and Hidden Markov Models (HMM)[8] for the recognition of emotions. The method classifies six fundamental emotions namely anger, dislike, fear, happiness, sadness and surprise from facial expressions and emotional speech. A bi-modal system was made on the basis of a rule-based classifier. It had an accuracy of 72%. However, the data extracted was inconclusive as 10% of data could not be sampled.

3. TEXT-BASED ANALYSIS

In the paper by Ramon, Mara, Gilberto that proposes a tutor system, an ASEM algorithm was utilized to recognize emotions from text input by a student. It showed a success rate of 80%. The input is a text line which is normalized. The words, numbers, and special characters with an accent are removed. The uppercase letters are converted to lowercase. Using the corpus Stop-words, Non-emotion words like he, she, the, etc. are removed.In corpus Semantic and improper words, the words are sought. Corpus New words is a set of words not found. The word features (PFA and emotion) are extracted when found in the Semantic corpus. The emotions are determined according to the word. Out of those emotions, the one with the greatest intensity is the output.

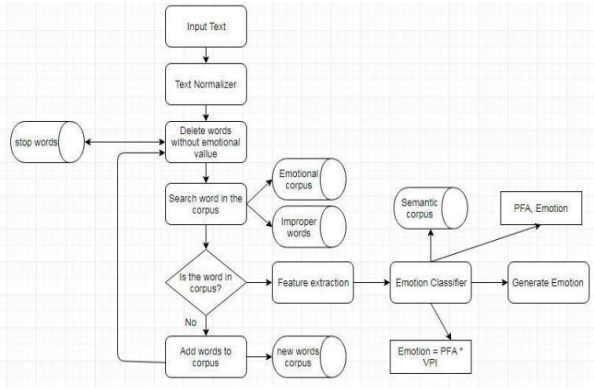

Figure 3.ASEM Algorithm(from Ramon, Mara, Gilberto, Computing Science 106 (2015)

In another paper[4], the authors extracted emotions from Youtube dataset, extracted transcripts and followed the Sentic computing paradigm developed by Cambria and his collaborators. It considers the text as expressing both semantics and sentics [6]. Concept level sentimental analysis and Concept level extraction are the two fundamental steps conducted. The text was represented using the bag of concepts feature. For each text, we extracted concepts using the concept extraction algorithm. Later, the concepts were searched in the EmoSenticSpace and if any concept was found then the corresponding 100-dimensional vector was extracted from the EmoSenticSpace. After that, the individual concept vectors were accumulated into one document vector using coordinate-wise summation. The concepts extracted from the text are assigned a polarity score. The polarity score is derived from the SenticNet. The summation of these scores gives us a scalar feature.

4. VIDEO-BASED ANALYSIS

Emotion detection is based on different expressions of face and these expressions are generated by variations in facial features. A video consist of varied facial frames that are beneficial in the matter of detecting emotions since a video is capable of capturing multiple facial frames and we could use appropriate methods to find the outputs. A broad study in emotion detection and analysis shows algorithms and techniques to capture facial images from a video that have been tested and concluded to be more accurate than still images. In a paper, authors have used Support Vector Machine to detect emotions from facial images and used PCA to extract the features and reduce the dimensions into 2-dimensional vector space. SVMs are memory efficient and effective in higher dimensional spaces. Also, OpenCV contains cascade classifiers in which Viola & Jones face detection algorithm is used. By using these classifiers the face region is detected from the image. It classifies the images into positive and negative images respectively. An image with a face is positive and without a face is negative. After classifying the training and testing images PCA is applied to the training set and classification into emotions Happy, Sad and Neutral is performed.

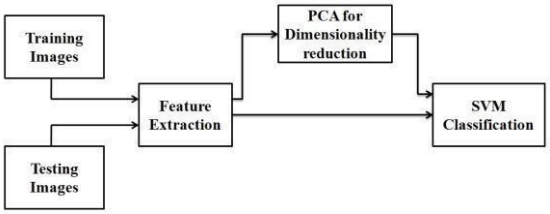

Figure 4.Block Diagram of image classification.

Generally, an emotion detection from video contains three stages - face detection, feature extraction and finally the classification stage. Some of the papers referred stated different methods to process images from the videos. Images are converted into 2-dimensional or 4-dimensional vector space. And commonly face detection algorithms studied were Viola-Jones Algorithm and Gabor filters are also applied on eyes and mouth regions to extract relevant features.

I. PROPOSED SYSTEM

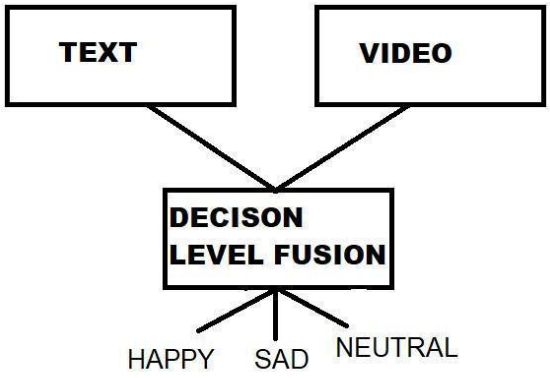

Figure 5.Proposed System

In this system, two modalities text and video are chosen as evident by work in the paper [1] this combination gives the best accuracy. The source for text is user's twitter account and source for video is real-time input while user answers a beck inventory depression based questionnaire.

Using Transfer Learning, the last layer is retrained for doing image recognition and labeling them as happy or sad. The inputs to the label classifier are the image user has uploaded, graph tokens coming from the trained graph and the two labels happy and sad. Thus this is how classifier is operated.

5. CONCLUSION

Thus a number of papers detailing the study of extracting features from various modalities, combining features from distinct sources and labeling user's current emotion is studied.

It can be inferred that combination of two or more modalities gives a better result than an individual modality. From the different bimodal systems, a combination of video and text gives best results. The Decision-level Fusion will be performed in such a way that preference will be given to the modality with more accuracy.

At Feature level, a summation of the various feature vectors is done and compared with the decision level feature vectors or scores. Then confusion matrix is generated to show deviation between actual output thought of by a person and this model's output. If the Text-based or image-based modality has the higher accuracy then the higher weight will be assigned to that particular one. Thus the combination or fusion is performed in this manner.

6. REFERENCES

[1]R. Zatarain-Cabada, M. Lucia Barrn-Estrada, J. Garca-Lizrraga, G.Muoz-Sandoval, "Java Tutoring System with Facial and Text Emotion Recognition", Instituto Tecnolgico de Culiacn,pp. 49–58"2015.

[2]C. Busso, Z. Deng , S. Yildirim, M. Bulut, C. Min Lee, A. Kazemzadeh, S. Lee, U. Neumann, S.Narayanan , "Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information", Proceedings of the 6th international conference on Multimodal interfaces,Pages 205-211,2004 .

[3]S. Emerich1, E. Lupu1, A. Apatean1, "Emotions recognition by speech and facial expression analysis" , Signal Processing Conference, 2009 17th European**,**2009.

[4]S. Poria, A. Hussain, E. Cambria, "Text-based sentiment analysis: Towards multimodal systems,"Proceedings of the 13th international conference on multimodal interfaces, Pages 169-176, 2011.

[5]L. Tian, D. Zheng, C. Zhu, "Image Classification Based on the Combination of Text Features and Visual Features", ISI Journal, Volume 28, Pages 242–256, 2013.

[6]L. Philippe Morency, R. Mihalcea, P. Doshi, "Multimodal Sentiment Analysis: Harvesting Opinions from the Web", ICMI '11 Proceedings of the 13th international conference on multimodal interfaces, Pages 169-176, 2011.

[7]P. Khorrami, T. Le Paine, K. Brady, C. Dagli, T.S. Huang1, "HOW DEEP NEURALNETWORKS CAN IMPROVE EMOTION RECOGNITION ON VIDEO DATA", Image Processing (ICIP), 2016 IEEE International Conference in 2016.

[8] R. Gupta, N. Malandrakis, and B. Xiao.," Multimodal prediction of affective dimensions and depression in human-computer interactions", Proceeding AVEC '14 Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Pages 33-40,2014.

[9]