HYBRID MODALITY LEVEL
CHALLENGES FACED IN MULTIMODAL
SENTIMENT ANALYSIS

V J Aiswaryadevi
aiswaryadevi.vj@skct.edu.in

S Soundarya
soundarya.s@skct.edu.in
Assistant Professor
Department of Computer Science and Engineering
Sri Krishna College of Technology
Coimbatore

Abstract- Aspects and sentiments expressed by human beings through behavior is computed systematically by neural networks and this approach have been tremendously appreciated throughout the research scientists nowadays. Multimodal sentiment analysis is the currently appreciated work in the deep learning field. Diverse modalities among image, videos, text, audio or fusion of the modalities here are through which human beings express their desirable information. Physically handicapped or the insecurity forces in the border of the nation can be the victims prone to be under vigilance monitoring for identifying the sentiment of actions they express. The paper provides the datasets available for various application areas under implementation and algorithms proposed till date to solve the application problems under implementation and followed by scientific challenges faced during fusion of two or modalities expressed here. The paper also includes the discussion about the implementation criteria for constructing the deep learning networks for various fusion levels of modality.

Keywords- Multimodal emotion recognition system, kernel ELM, ANP, Sentiment Analysis, MKL

1. INTRODUCTION

Pattern analysis and machine intelligence makes use of HCI tools to promote points on face for capturing emotions using camera as the primary mean [1]. ImageNet and Flickr images are used as image data set to promote two modalities (text and ANPs and image) with the help of deep multimodal sentiment model by having images and ANPs pre trained in the first layer followed by the recurrent network carrying out the process of mapping word tag to the emotions in the high dimensional space. This layer is followed by the softmax output layer preferably for the list of derived word tags from the images [3]. Additionally [3] has implemented the deep learning for emotion word tags on tumblr data set and classified the images on “happy”, “calm” and so on for about 15 different emotions. Visual and acoustics. Older technologies were making use of Facial Action Coding System (FACS) [5]. An attention network scheme in deep learning networks is used for focusing the input features alone excluding the unwanted background images. Linguistic, acoustic and visual modalities are combined various level and video summarization is also adopted by researchers with the help of complex algorithms. [17] focuses on approaching the linguistic features using SVM on frame wise ASR data in YouTube videos and audio-visual data with the help of BLSTM model. Finally the results are fused to derive the multimodal sentiment analysis model for the combination of audio, visual and text data. High level of accuracy is derived from the model discussed here. Fusion of LSTM and RNN algorithm is also used for emotion recognition in [18]. Additionally Adadelta algorithm was required to reduce the squared loss in the prediction results.
2. **DATA SETS AND TOOLS**

Utterance level word tags are used to construct the Multimodal Opinion Utterances Data set (MOUD) [7]. Li, W., (Li, W.. 2011) implemented the multimodal SA using linguistic, acoustic and visual analysis at utterance level. Each utterance denotes the starting and ending point of time for an acoustic information. Annotations are done by elan tool which is widely known for video and audio annotator. Each and every instance of audio chat signal is annotated by elan and the video is segmented and annotated at the same pace by elan. Weihong Li used MOUD for multi objective optimization on face recognition with the 2D Gabor Gaussian function for image compression and LDA for feature selection and extraction and finally SVM for MMSA.

Kernel ELM is used for better utterance level classification [8]. Wang., (Wang., 2017) has implemented the HMM model for acoustic learning at the utterance level and Kernel Extreme Learning Machines were used in the Deep Neural Networks (DNN) for emotion recognition. Softmax function on the DNN layer was replaced by the kernel ELM for 90.56% to 92.72% of weighted accuracy in DNN-ELM approach on the emotion recognition task.

Kernel ELM is used for better utterance level classification [8]. Wang., (Wang., 2017) has implemented the HMM model for acoustic learning at the utterance level and Kernel Extreme Learning Machines were used in the Deep Neural Networks (DNN) for emotion recognition. Softmax function on the DNN layer was replaced by the kernel ELM for 90.56% to 92.72% of weighted accuracy in DNN-ELM approach on the emotion recognition task.

![Image](image.jpg)

**Fig. 1. Examples of Image with Positive Tweet from SentiBank Twitter data set**

Twitter short messages or microblogging content is even used as dataset for text sentiment analysis. It contains up to 140 words in a tweet. Twitter messages with tweets or short comments are used as SentiBank dataset for many researchers to do sentiment analysis. Public sentiment about issues and happenings in trend is vastly available in the twitter tweets. The training data taken from SentiBank exceeds nearly 17000 tweets. SentiBank was proposed by Borth Et al[11]. OpenEAR is a tool used in [21] to classify the multimodal sentiment expressed through spanish videos with 2-8 minutes length by converting them into linguistic features.

![Image](image2.jpg)

**Fig. 2 Online Sentiment Annotation Interface**

- MOSI—contains video-based topic reviews annotated by sentiment polarity
- 93 videos, 2199 utterances
- MOUD—contains product review videos annotated by sentiment polarity
- 498 Spanish utterances, 55 individuals
- IEMOCAP—contains scripted affect-related utterances annotated by emotion categories
- Youtube - 47 videos, 280 utterances.
- Transcription and segmentation on videos can be performed using Transcriber software and sentiment annotations performed using elan software. Acoustic features can be extracted using openEAR and visual features extracted using CERT (Computer Expression Recognition Toolbox).

3. **FUSION LEVEL**

3.1 **Image and Text Fusion**

Bilateral Correspondence model (BC-LDA) for image and words was adopted by Baecchi.,(Baecchi., 2016). BC-LDA incorporated
neural network based models such as Skip-gram and Denoising autoencoders for sentiment analysis on microblogging content such as Twitter short messages. Social Images with tweets and corresponding tags are analyzed for sentiment polarity with Multinomial Naïve Bayes algorithm. Supervised NLP task is implemented on the neural network with the help of hierarchical softmax function. PCA or LSTM is used as the autoencoder hidden layers implementation to discover lot of hidden features.

Fig 2. Denoising Autoencoders

Sliding window framework was used for the image analysis with tweets on neural network. It is also used to ignore the noisy labels or lost labels. Positive, Negative, Neutral, and unsure tweets are obtained as output in sentiment analysis process in a sliding five-minute window. An online annotation tool was invented by Wang., (Wang., 2012) in [10]. The users of online annotation tool are allowed to submit their own opinion of the tweet. 59% of accuracy was achieved by [10] on the four category such as positive, negative, neutral and unsure tweets. By combining the softmax and euclidean loss in the deep learning output layer, 78.89% of accuracy was obtained. Whereas Softmax function is used for sentiment polarity variation and Euclidean function is used for sentiment strength variation.

3.2 Text, Image and Audio Fusion

Attention scheme embellished with memory networks in the deep learning architecture is adopted for memory under sentiment analysis using Progressive Convolutional Neural Network algorithm (PCNN). PCNN was proposed by [15] in the year 2015 high performance attention scheme for autoencoders. Felix 2017 made use of Progressively Trained Attention Shift Convolutional Neural Networks (PTAS-CNN) and Deep Convolutional Generative Adversarial Networks (DCGAN) for predicting gender from facial recognition on human facial images. AlexNet architecture is used for progressive CNN training. AlexNet has 60 million parameters and 650,000 neuron consisting of 5 convolution layers and 3 fully connected layers with a final 1000-way softmax. Each training data is trained with 1000 epochs and with a batch size of 128.

Wolmer.,[17] in 2013 extracted the BoW from the linguistic features and classified the same using SVM. The combination of results gained from SVM on linguistic features and BLSTM on audio visual data are taken into fusion for high accuracy prediction on youtube videos. Chao., [18] used the LSTM-RNN hybrid fusion model to obtain sentiment prediction on RECOLA dataset, an electro-cardiogram (ECG) and electro-dermal activity modalities. Similarly CNN was combined with LSTM for FaceScrub data set which would also require the loss reduction algorithms to be implemented with it.

4. DISCUSSION & RESULTS

Currently linguistic features, acoustic and visual features are taken into consideration for fusion with accuracy obtained as 86%, 79%, 67.31% of accuracy to the maximum respectively. [17] developed a sentiment analysis model on linguistic features using SVM and used Bidirectional LSTM algorithm for extracting the sentiment from audio and visual data. The predicted results of hybrid algorithms fusion are derived from the output of SVM and BLSTM algorithm which was implemented for youtube videos.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>BC-LDA</th>
<th>BLSTM + SVM</th>
<th>LSTM-RNN</th>
<th>CNN</th>
<th>PCNN</th>
<th>PTAS-CNN</th>
<th>CRMKL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio</td>
<td>78.89%</td>
<td>86.61%</td>
<td>78%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Audio    | 86.2   |
CNN was used as a trainable feature extractor to extract features from textual data. 7 layers were used in CNN proposed by [19]. The seven layers were as follows: Input layer with 211,114 neurons followed by convolution layer with kernel size 3, 50 feature maps followed by Max-pool layer with max-pool size 2 followed by convolution layer with kernel size 2 and 100 feature maps followed by max-pool layer of size 2 connected with a fully connected layer of 500 neurons finally ending with output softmax layer of 2 neurons defining the polarity values as positive or negative.

CNN is trained by standard backpropagation procedure. Whereas recent trends have proven 95% accuracy in multimodal sentiment classification on three modalities with the help of Convolutional recurrent multiple kernel learning architecture proposed by [22].

CONCLUSION
As per the discussion, deep learning networks are nowadays used to replace the feature extraction procedure and train the system to classify the input modalities derived with the positive or negative polarity. Multi-Kernel learning (MKL) is employed with these deep learning algorithms to enhance the classification procedure on multiple modalities. Extreme Learning Machines are used to enhance the learning rate of features from noisy input modalities. It is observed from the results in Table. 1 that CNN with multiple kernel learning is used to extract the highest accuracy in classifying the sentiment polarity from the multimodal input data fusion. 85.67% of accuracy is reached without English translation by Google translator from Spanish language. It is preferred to construct the deep learning network with denoising encoders inculcated with MKL in multiple modalities. By increasing the neurons in Max pool layer feature extraction can be effectively increased in accuracy with each neuron trained with probabilistic multi-kernel learning. Convolutional Recurrent MKL (CRMKL) is used to combine the sentiment features from multimodal inputs listed as commonly audio, video, text.

REFERENCES


utterance-level multimodal sentiment analysis. In Proceedings of the
20. 2015 conference on empirical methods in natural language processing (pp. 2539-2544).
31. Priyanka, G. PREDICTION OF AIRLINE DELAYS USING K NEAREST NEIGHBOR ALGORITHM.


