Sentiment analysis by deep learning approaches

Sreevidya P.¹, O. V. Ramana Murthy², S. Veni³

¹Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, India
²Department of Electrical and Electronics Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, India

ABSTRACT

We propose a model for carrying out deep learning based multimodal sentiment analysis. The MOUD dataset is taken for experimentation purposes. We developed two parallel text based and audio based models and further, fused these heterogeneous feature maps taken from intermediate layers to complete the architecture. Performance measures—Accuracy, precision, recall and F1-score—are observed to outperform the existing models.

Keywords:
Bimodal
CNN layers
MOUD
Multimodal
Word embeddings

This is an open access article under the CC BY-SA license.

Corresponding Author:
Sreevidya P.,
Department of Electronics and Communication Engineering,
Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Coimbatore, 641112, India.
Email: sreevidyapmenon@mail.com

1. INTRODUCTION

Great efforts are needed to develop machines that can mimic the natural ability of human beings to understand emotions, analyze situations and understand the sentiments associated with the context. The sentiment analysis is an effective mechanism to explore the socio-economic or demographic influence in human reciprocation. With the availability of a plethora of opinionated videos in social media, multimodal approaches in the sentiment analysis is gaining attention. Opinionated videos are highly unstructured; hence verbal and non-verbal cues are complementary in the sentiment analysis at this juncture. That means analyzing the communication in audio, visual along with text modalities has to be incorporated for achieving effective solutions. Most of the existing frameworks for classifying the sentiments are based on transcriptions based analysis [1] and the use of lexicons, but not much of the literature mines through the vocal and visual cues embedded in the videos. The voiced communication can give more information regarding the human empathetic conditions [2]. This work aims in fusing information from different modalities for the sentiment analysis.

The primary benefit of analyzing videos along with texts is that the rich set of behavioral cues present in audio and video recordings can yield enhanced models. The vocal modulations, facial expressions and gestures in the visual data, along with textual data, help to analyze the affective domain of the opinion holder in a better way. Thus, a combined text, vocal and visual data help to create a more robust and emotion specific sentiment analysis model [3]. There is an array of techniques available for carrying out
the sentiment analysis, through incorporating machine learning and deep learning paradigms. There are multi-faceted challenges associated with extracting information from different modalities and to fuse them together for the analysis. We propose a bimodal approach for predicting the sentiments using deep learning based techniques.

The proposed deep sentiment analysis framework includes:

a. A Convolutional Neural Network (CNN) based model with max-pooling, and dense layers to process features extracted from sentence level utterances.

b. A model for processing transcriptions which is trained with CNN layers. The sentence level text is mapped into a vector space using a word representation learned by word embedding.

c. A fusion model containing the features extracted from specific layers of both audio and transcription models.

Conventionally, the problem of sentiment analysis is based on textual information. The analysis is carried out at word level, sentence level or document level. Pre-processing steps include cleaning of texts, removal of white spaces, expanding the abbreviations, stemming, removal of stop words, negation handling followed by feature selection and finally classification techniques [4]. The classification techniques can be divided into machine learning (ML) based approaches and lexicon based approaches. The ML based supervised learning approaches include probabilistic models such as Naive Bayes classifiers [5] or Bayesian classifiers [6]. Because of the sparse nature of the text data, the Support Vector Machines (SVMs) are effectively used for classifying transcription sentiments, both for multi-class and binary class problems. Li and Li [7] used SVM for classifying sentiments in micro blogs. Neural network and SVM were applied for sentiment analysis and compared by Moraes et al. [8].

The automated lexicon based approaches are split into dictionary based approaches and corpus based approaches [9]. The dictionary based approaches focus on finding the opinion seed word, whereas corpus based approach begins with a seed list of opinion words. The corpus based approach is limited due to the difficulty in preparing huge corpus and normally employs either statistical based techniques [10] or semantic based techniques [11]. With the increased presence of multimedia tools, especially on social media platforms, sentiment analysis could not be restricted to transcription based analysis. This has paved ways to multi-modal approaches in sentiment analysis. While the unimodal text based analysis was focused at text pre-processing and selecting suitable methods for analysis, there were greater challenges in multimodal approaches. In conventional analysis, rule based methods using lexicons and data driven methods using large, annotated databases [12, 13] are popular. But in multimodal analysis, the heterogeneous dimensions from image, text and audio signals are to be combined together. There are three strategies popular for multimodal fusion, viz, early fusion, late fusion and intermittent fusion. The work in [14] apply early fusion of low level and mid level features extracted from human faces to have group level emotion detection. A major shortcoming of early fusion technique is the absence of detailed modeling for view-specific dynamics, which will affect the modeling of inter-view dynamics which causes overfitting of input data and models based on late fusion are normally good in modeling view-specific dynamics. Late fusions have shortcomings in modeling the cross-view dynamics since these cross-modality dynamics are considered to be more difficult [15]. The traditional hand crafted feature extraction methods paved ways to deep learning techniques, additionally, the Recurrent Neural Networks (RNN) and Long Short time Memory (LSTM) could take up the spatial and temporal information directly from the raw data [16].

2. RESEARCH METHOD

A bimodal approach with utterances taken in audio and text formats is proposed here for sentiment analysis. The MOUD dataset containing opinionated utterances in sentence level [13] is taken for experiments. The architecture developed is shown in Figure 1. Utterances audio and text are the inputs of the framework and the output is binary classification-positive or negative polarity. The architectural pipeline includes two parallel independent deep learning framework having unimodal processing of audio and text utterances. The deep neural features extracted from these individual modalities are fused together and given as input to the final CNN layers to apply the bimodal fusion.

2.1. Unimodal approaches

The proposed system intends to develop individual models for transcriptions and audio signals at the first stage. Later, a bimodal architecture is developed by integrating the independent models. Each stage is described as follows.
2.2. Audio features

Analyzing the speech as sound will help the system to focus on classifying the polarity of the sentence either as positive or negative by eliminating the language barrier. As for the audio utterances are concerned, the audio features are extracted from the input audio signal by the application of a third party acoustic feature extraction tool called OpenEar [17, 18]. The features extracted are using SMILE feature extractor and Low Level Descriptors (LLDs) including 13 Mel-/Bark-Frequency-Cepstral Coefficients (MFCC) which typically ranges between 300Hz to 5KHz, prosody, energy, voice probabilities and spectral coefficients resulting in a feature vector set 27 for each utterances. The features are extracted with a frame sample rate of 25ms and standardization is performed for speaker normalization.

This feature set is applied to a deep learning framework starting with a convolution layer including 256 filters of size three. The convolution layers are interleaved with a max-pooling layer so that the filter output size is reduced by a factor of two. The network goes deeper in this fashion by implementing convolutional layers of size 3 with the number of filters as 128 and 64 respectively. After the convolutional operations, three consecutive dense layers are added to flatten the network and gradually reduce the output. The max-pooling layer will reduce the dimensions of the set of feature by a factor of 2. Next, a dropout is applied as a regularization technique in order to reduce the number of network connections which helps to avoid the overfitting of the network layers just before flattening the layers. The non-linear Rectified Linear Unit (ReLU) is applied as the activation function for the hidden layers and final decision on type of the sentiment is based on the output of the softmax function.

The transcribed utterances in the MOUD dataset, which is annotated is combined into a single CSV file. Initially the database undergoes certain pre-processing steps so as to avoid the outliers. Subsequently, the data is given to a tokenizer to create the vocabulary. The word embeddings are used to get the word vectors. It is like the concatenation of words. This feature sets are trained in a deep neural framework II to carry out the output the sentiments classification.

2.3. Textual features

Primarily, text data must be encoded as vectors before applying it to the deep learning model. For hat (i) sentences are pre-processed and tokenized to get the integer representation. The start and stop words as well as the wild characters are removed during pre-processing. At the same time all the words are onverted to lowercase letters. Keras Tokenization API is used for tokenizing the sentences. (ii) Finally, The word embeddings are applied to convert the positive integers to dense vectors of fixed size. The dense vectors represent the projection of the words into a continuous vector space whereby each word will have a unique vector representation. As a result, the words will be in a coordinate system, where, related words based on the corpus relationships will be placed close to each other. The vector values are learned in a way.
that resembles to the method of learning in a typical neural network [19]. The feature vectors obtained are padded to a window of standard length of 60. This standardized vectors are given as input of the deep learning model. The first input was given to the convolutional layer of size 3, consisting of 128 filters, followed by a global max pooling layer. The convolutional layer systematically apply learned filters to the input data so as to create feature maps that summarize the presence of the strong feature set in the input data. The global max-pooling layer will down sample each feature map into a single value which is the maximum value of the patches of the feature set [20]. In this way the problems due to overfitting of the fully connected layers can be minimized. Subsequently, there are two dense layers. All layers except the final dense layer is with ReLU activation function, whereas the final decision making layer has softmax as the activation function. The model has 96,796 parameters to be learned during the training. Typically, if there are n words in a sentence [21], it can be tokenized as an integer vector T, where

\[ T = \{T_1, T_2, ..., T_n\}, \quad n = 60 \]  

(1)

\[ T \in \mathbb{R}^{1 \times d} \text{ dimension, } d \text{ denotes the word length. By applying the word embeddings each token will be vectorized consisting of the feature representations of the required transcriptions. It is given as embeddings,} \]

\[ E = \{W_e, T\} \]  

(2)

where \( W_e \) is the parameters to be tuned and \( W_e \in \mathbb{R}^{d \times |T|} \). The hidden layer output is represented as

\[ h_i = f(E, \theta_i) \]  

(3)

where \( \theta_i \) is the weights and bias parameters and the final activation layer is the softmax layer [22]. For a given class \( h_i \) the softmax function is represented as:

\[ h_i = \frac{e^{h_i}}{\sum_{j=1}^{C} e^{h_j}} \]  

(4)

where \( h_j \) are the values inferred by the net for each class in C.

2.4. Bimodal Framework

In the proposed model, individual, parallel networks were trained initially. Later, the intermittent layers of both these networks are extracted as feature input for the bimodal framework. In this way, the complementary information from both the modalities could be taken effectively. The 3rd layer of the textual model and 6th layer of the audio model are optimally selected and extractadas features for the final fusion model. The global max-pooling layer in text modality significantly reduced the size of the feature map and the same was done in audio modality thorough downsampling the dense layer. Features from these two layers are concatenated and it is applied as input to the third combined model. The feature sets are applied directly without any pre-processing. This model is also a deep neural network consisting of convolutional layers and max-pooling layers. The output from the model will classify the utterances as positive or negative polarity. The decision vector formed by combining the text and audio modalities are improving the performance of sentimental analysis considerably compared to individual modalities alone. The final decision on sentiment classification is taken based on the softmax activation function.

The experiments are conducted on MOUD dataset both on individual and combined modalities. During the training phase of the proposed model, the weights are adjusted to minimize the loss function. The hyper-parameters of the proposed neural network model are tuned fusing with the weights to further optimize the results. The role optimizer in deep neural network models are to minimize the cost function \( J(\theta) \), where the parameter \( \theta \in \mathbb{R} \), with respect to the performance measure P, under consideration by applying gradient descent algorithms. The cost function can be represented as:

\[ J(\theta) = E_{x,y} - \tilde{P}_{data} \log [P_{mod,\theta}(x, y; \theta)] \]  

(5)
A faster convergence of the model is achieved by selecting a proper learning rate as in:

$$\theta = \theta - \eta_0 \nabla J(\theta)$$  \hspace{1cm} (6)

where $\eta_0$ represents the learning rate of the gradient descent algorithm. The optimizer algorithm we used for comparison purpose are stochastic gradient descent algorithm(SGD), Root Mean Squared prop(RMSProp) and ADAptive Moment estimation(ADAM). The SGD does the parameter updates for all the training examples in the training set with a prefixed learning rate[23]. In the RMSProp algorithm proposed by Geoffrey Hinton, instead of letting all of the gradients to accumulate the value of its momentum, RMSProp algorithm only accumulates gradients in a fixed window. Adam optimizer computes adaptive learning rates for each parameter considered in the algorithm and it stores the exponentially decaying average of the square of the gradients of the previous values[24].

2.5. The MOUD Dataset

The Multimodal Utterance Opinion Database (MOUD) introduced by Perez et. al. [13] is an opinionated dataset in Spanish language. It consists of product review and recommendations in utterance level from 80 speakers collected through YouTube videos. From the available 498 videos we selected 438 recordings for our work, which showed consistency among speech and text modalities and on an average, each one of the video has 6 utterances of 5 seconds duration with a standard deviation of 1.2 seconds. The contents of each one of the video clips were transcribed through manually processing the verbal statements for its connotations. Annotations of the dataset was done using Elan tool for sentiment analysis. Both audio and video modes are annotated using the tool. Two annotators independently annotated the polarity of the utterances as positive, negative or neutral. In our classification problem, positive and negative sentiments were only considered.

3. RESULTS AND ANALYSIS

The objective is to classify the sentiments in the videos based on the polarity as positive or negative through analyzing the MOUD dataset. A combined audio and text model was developed by implementing deep neural networks. The dataset was optimally divided into a train-test ratio of 80:20 for developing the model and testing the data. The categorical cross entropy, which is a combination of softmax and cross entropy function was taken as the loss function for training the model. The unimodal features are applied to the two parallel subnets and the outputs of from intermediate hidden layers are optimally selected. These selected values are fused to get and the same will be acting as the input to the final subnet. Several experiments were conducted before fixing the proposed architecture. We compiled the model with different hyper-parameters also. There was some significant changes based on the optimizer selection. The minibatches can offer the effect of regularization. The minibatch selected was 32 for the proposed model. There were significant changes in the performance of the model based on the optimizer selection. The output of the proposed system is one hot encoded. The results of the experiments are tabulated in Table 1. The performance of the proposed model was evaluated using different performance matrices viz, accuracy, precision, recall and F1-score.

<table>
<thead>
<tr>
<th>Mode</th>
<th>optimizer</th>
<th>accuracy</th>
<th>precision</th>
<th>recall</th>
<th>F-1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio</td>
<td>Adam</td>
<td>0.76</td>
<td>0.8</td>
<td>0.71</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>0.72</td>
<td>0.71</td>
<td>0.75</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>Rmsprop</td>
<td>0.7</td>
<td>0.68</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>text</td>
<td>Adam</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>0.73</td>
<td>0.74</td>
<td>0.6</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Rmsprop</td>
<td>0.72</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>audio + text</td>
<td>Adam</td>
<td>0.84</td>
<td>0.86</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>SGD</td>
<td>0.75</td>
<td>0.86</td>
<td>0.6</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Rmsprop</td>
<td>0.72</td>
<td>0.75</td>
<td>0.68</td>
<td>0.72</td>
</tr>
</tbody>
</table>

The graphical representation of accuracy on training each epoch is shown in Figures 2-10. The effects of different optimizers are highlighted here. The ADAM optimizer is showing the optimal results. The SGD is giving fluctuations during convergence. The parameters selected for SGD are as learning rate=0.001 and momentum = 0.9. In the case of RMSProp algorithm, the values are also the same. For Adam optimizer, in addition to the above values, the decay rates are also fixed as 0.999.
Further, we compared the performance of our algorithm with some of the existing algorithms and the superiority of the combined audio and text proposed architecture is quite evident from Table 2. The proposed model was compared with four of the existing state of the art methods. Poria et. al., [25] proposed a speaker exclusive technique for analyzing the sentiments embedded in the utterances. Wang et. al., [26] proposed to mitigate the problem of generalizability to a larger margin. Poria et. al., [27] proposed the Convolutional Recurrent Multi Kernel Learning (CRMKL) model using CNN networks exclusively for training the model, and the combined network takes the best features only by Principal Component Analysis (PCA) and they used SVM for the decision making.

The results of the experiments were tabulated in Table 2. It shows the test results with and without feature selection. Our results and the results obtained by Poria and his team are compatible. Cambria et. al.,
[28] presented a deep learning architecture focusing on speaker independent systems. Our method performed much better than this proposed work. Tsai, et. al., [29] proposed a multimodal factorized modal (MFM) with multimodal discriminative and modality-specific generative factors.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>48.4</td>
<td>52.2</td>
<td>74.5</td>
<td>53.7</td>
<td>71.1</td>
</tr>
<tr>
<td>Audio</td>
<td>53.7</td>
<td>54.4</td>
<td>79.8</td>
<td>53.7</td>
<td>76.0</td>
</tr>
<tr>
<td>Text + Audio</td>
<td>57.1</td>
<td>57.4</td>
<td>83.8</td>
<td>57.1</td>
<td>84.1</td>
</tr>
</tbody>
</table>

Sentiment prediction results on MOUD are depicted in Table 2. The best results are highlighted in bold and SOTA shows the changes in performance over previous state of the art (SOTA) results. The improvements are highlighted in bold in Table 3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Majority</td>
<td>60.4</td>
<td>45.5</td>
</tr>
<tr>
<td>RF</td>
<td>64.2</td>
<td>63.3</td>
</tr>
<tr>
<td>SVM-MD</td>
<td>59.4</td>
<td>45.5</td>
</tr>
<tr>
<td>THMM</td>
<td>61.3</td>
<td>57.0</td>
</tr>
<tr>
<td>EF-HCRF</td>
<td>54.7</td>
<td>54.7</td>
</tr>
<tr>
<td>EF-LDHCRF</td>
<td>52.8</td>
<td>49.3</td>
</tr>
<tr>
<td>MV-HCRF</td>
<td>60.4</td>
<td>45.5</td>
</tr>
<tr>
<td>MV-LDHCRF</td>
<td>53.8</td>
<td>46.9</td>
</tr>
<tr>
<td>CMV-HCRF</td>
<td>60.4</td>
<td>45.5</td>
</tr>
<tr>
<td>CMV-LDHCRF</td>
<td>53.8</td>
<td>47.8</td>
</tr>
<tr>
<td>DF</td>
<td>67.0</td>
<td>67.1</td>
</tr>
<tr>
<td>EF-LSTM</td>
<td>67.0</td>
<td>64.3</td>
</tr>
<tr>
<td>EF-SLSTM</td>
<td>56.6</td>
<td>51.4</td>
</tr>
<tr>
<td>EF-BLSTM</td>
<td>58.5</td>
<td>58.9</td>
</tr>
<tr>
<td>EF-SBLSTM</td>
<td>63.2</td>
<td>63.3</td>
</tr>
<tr>
<td>MV-LSTM</td>
<td>57.6</td>
<td>48.2</td>
</tr>
<tr>
<td>BC-LSTM</td>
<td>72.6</td>
<td>72.9</td>
</tr>
<tr>
<td>TFN</td>
<td>63.2</td>
<td>61.7</td>
</tr>
<tr>
<td>MARN</td>
<td>81.1</td>
<td>81.2</td>
</tr>
<tr>
<td>MFN</td>
<td>81.1</td>
<td>80.4</td>
</tr>
<tr>
<td>MFM</td>
<td>82.1</td>
<td>81.7</td>
</tr>
<tr>
<td>Proposed Method</td>
<td>84.1</td>
<td>84.1</td>
</tr>
<tr>
<td>SOTA</td>
<td></td>
<td>2.4</td>
</tr>
</tbody>
</table>

4. CONCLUSION

An analysis of the existing methodologies for sentiment analysis and the comparison with the proposed bimodal sentiment analysis system is carried out here. The proposed framework establishes the superiority of bimodal approaches over unimodal approaches. We are incorporating the the powerful CNN based deep learning techniques for the test case. The intermediate level feature fusion method is adapted here. The sequential and correlated information is collected through word embeddings in the textual data and audio feature extractions. For further analysis, so as to increase the accuracy of the performance of the model non verbal communications like jesters and images can be incorporated. The multi-modal approach can integrate all the information related to the communication, which in turn can make the human computer interactions more realistic and meaningful.

REFERENCES


**BIOGRAPHIES OF AUTHORS**

Sreevidya P. is currently a Research Scholar in the department of Electronics and Instrumentation at Amrita Viswa Vidyapeetham, Coimbatore. She was graduated from Calicut University and did her Masters at PSG college of Technology, Coimbatore. She has got more than 10 publications in her credit. Her major areas of interest are artificial intelligence, affective computing, multimodality and image processing.
Dr. O. V. Ramana Murthy currently serves as Assistant Professor (SG) at the Department of Electrical and Electronics Engineering, School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore Campus, India. He has more than 9 years of research experience in reputed Universities including National University of Singapore, Nanyang Technological University, Singapore and University of Canberra, Australia. He has more than 20 publications in major international journals and conferences to his credit. He has participated in a number of workshops on Big data, deep learning, statistics.

Dr. S. Veni joined Amrita Vishwa Vidyapeetham as a faculty in the year 2001 under the Department of Electronics and Communication Engineering. She is currently working as Associate Professor. She received her AMIE degree from Institution of Engineers, Calcutta, in the year 1994 and M. E. degree (Applied Electronics) from Bharathiar University, Coimbatore, in the year 1998. She obtained her Ph. D. degree in the area of Image Processing from Amrita Vishwa Vidyapeetham in January 2012. Her areas of interest include Signal and Image Processing, Hardware Implementation of Signal and Image Processing Algorithms. She has published nearly 50 papers in international journals and conferences. She has received best paper award four times for the papers she had presented in the conferences. She is an Associate Member in the Institution of Engineers, Member in ISTE, Member in IETE, BOS member, Doctoral Committee Member (Amrita Vishwa Vidyapeetham, Anna University, Karunya University and Kerala Agricultural University), Academic Community Member in International Congress for global Science and Technology (ICGST). She has served as a reviewer and session chair in the International conferences, reviewer in reputed journals like Inderscience and IET Image Processing.