CNN based framework for Intelligent Multimodal Sentiment Analysis

Mrs. Raviya .K[1] , Dr. Mary Vennila .S[2]

[1]Research Scholar, PG & Research Department of Computer Science,
Presidency College, Chennai. Email: raviya.mca@gmail.com

[2]Associate Professor, Head & Research Supervisor, PG & Research Department of Computer Science,
Presidency College, Chennai. Email: vennilarhymend@yahoo.co.in

Abstract
Sentiment analysis of online user created text content has a very prominent role for many social media analytics tasks. In today's world, additional images and videos are extensively being used by the social media users for sharing their opinions and experiences. Sentiment analysis of huge scale in text and visual data mainly supports to better extract user sentiments toward brands or topics. To counterbalance with the growth of enormous multimodal data, there arises the immediate necessity to bring out an intelligent multi-modal sentiment analysis framework efficiently mine information from multiple modalities. Prior studies focused mainly on single modality content such as text or image. Here, we plan to achieve a new frame work for multi-modal sentiment analysis using CNN based feature extraction from various modalities. The multi model sentiment analysis on two publically available datasets shows the intelligent of our models and we gain a consistent performance improvement overstated by intermingle text, audio and visual features.

Keywords: *Sentiment analysis, multimodal data, CNN, SVM*

I. Introduction
Views and viewpoints have an incredible role in decision making. Social network sites, blogs, forums, e commerce websites, today, have given internet user with platform to put forth their ideas, views. This has resulted in, the modality of massive social media data as a wide platform and not to the single text mode. In microblog sites, for example, more and more users are inclined to put up multimodal tweets, adding an image in their tweets, which brings new challenges to social media analytics in handing massive quantity of multi-modal scale social media data. Analyzing user sentiments on two or three input modes improvise the accurateness of the analysis, in contrast to the traditional text-based sentiment analysis. The main improvement of analyzing video provides multi-modal facts in terms of spoken and visual modalities. The voice modulations and face movements in the visual data, along with textual data, provide true views about anything by opinion holder. For multi-modal sentiment analysis Figure-1 shows, how each modality bestows to the identification of sentiments.

Recently, many connected work has been projected to deal with multi-modal sentiment analysis, and deep neural network-based ways show superiority in performance. One main challenge in multi-modal sentiment analysis task is the fusion of all modalities of samples, including text, image, video or speech. Even though related deep neural network-based method helps to create a novel multimode sentiment analysis model. Thus multi-modal sentiment analysis is a projecting research area in recent years particularly within the contextof social media huge information.

Earlier related studies mainly targeted on single mode sentiment analysis especially on text. Today, besides the traditional machine learning-based methods, deep neural network has gained increasing attention in extracting textual representation.

Meanwhile, inspired by their superior performance in image classification, CNNs has also been used in image sentiment analysis. To subdue the limitations in the previous work,This article suggest anend-to-end framework for large-scale multimodal sentiment analysis based on CNN.

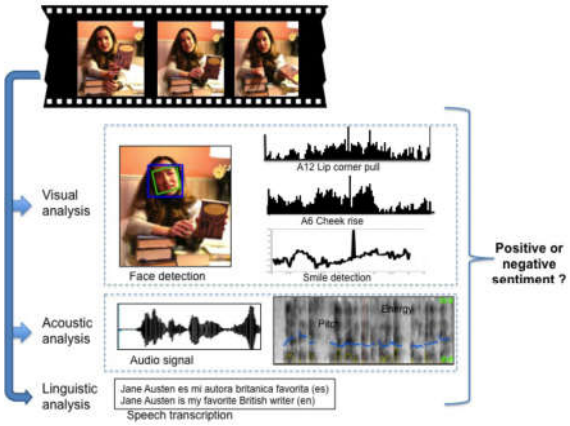


Figure I : Multimodal feature extraction

II. Applications
Sentiment Analysis has been extensively used for understanding the nature of a context. Few areas where It can be applied are

1. Businesses and organizations

Much of the business strategies are being escorted with respect to the response from the customers. Companies intend to satisfy the needs and demands of the users, thus strategic moves of companies are driventhrough public opinions and views. With the world connected through technology events havea global influence; theissue/failure on one place of the nation has an impact on the other part of the country. So it has become inevitable to drive products/services according to the public viewpoint. Nowadays Business men are investing much amount of money to find the users sentiment.

2.Individual products analysis and decision making

With the support of sentiment analysis it has become easier to examine different products and make the choices accordingly. This kind of analysis also helps to select a product based on its feature specifications. The comparison between two products has also been made quite easier. Decision making is a vital part of our life. It scales from which products to buy, which restaurant to go, to which bank insurance policies to go for and which investments to make. People used to decide and select from the available options based on the general opinions expressed by other users.

3. Ads placements

Display an ad when one appreciates a product and place an ad from a participant if one analyzes a product. Whenever one express opinion about a product by analyzing that review we can say whether it is positive or negative.

4.Recommendation systems

Most of the websites we visit have a recommendation system in-built to assist us, ranging from sites related to books, online-media, entertainment, music, film industry to other forms of art. These systems use our personal information, previous history, likes and dislikes and our friends' information to make suggestions.

5.Designing and building innovative products

When open to tough competition and open to critics through public reviews and opinions, sentiment analysis leads to better analysis of the product usability and human-friendly nature. It creates an environment for better and more innovativeproducts.They can get Associate in Nursing user satisfaction regarding themerchandise from the quantitative relation of positive to negative tweets.

6.Computing customer satisfaction metrics

They can get an user satisfaction about the product quantitative relation of positive to negative tweets.

7. Identifying enemies and sponsors

It is useful for customer support service, by noticing unhappiness or problems with the products. It also helps to find who are cheerful with the brand or facilities and their familiarities are utilized to promote our products.

III. Convolutional neural Network

In machine learning, The CNN is a category of deep, feed forward artificial neural-network most commonly applied in computer vision. Now it is used in NLP for text classification. CNN architecture comes in several variations. However In general It consists of convolutional and pooling layers which were grouped in to modules. We first fine-tune this network as a machine learning algorithm to serve as a feature extractors for visual and text sentiment analysis. In our system, we applied this new method to multi-modal sentiment analysis using deep neural networks such as CNN by combining visual recognition and NLP.The proposed system with CNN achieved good validation accuracy with high consistency.

IV. Related work

Sentiment analysis is a focused area of study in the past decade, particularly on twitter textual data, e.g. early work by Pak &Paroubek (2010) showed that emoticons could be used to collect a labeled dataset for sentiment analysis. Visual sentiment analysis is totally different from text analysis as it requires a higher level of abstraction to grasp the message carried out by an image (Joshi et al., 2011). Borth et al. (2013) pioneered sentiment analysis on visual content with SentiBank, a system extracting mid-level linguistics attributes from images. These linguistics features are the classifiers output that may guesst the significance of an image with regard to one of the emotions inside the Plutchik's wheel of emotions (Plutchik, 2001). Motivated via the development of deep learning strategies, You et al. (2015) used convolutional neural networks on Flickr and Twitter for binary sentiment classification. However, research concerning image annotation confirmed that combining text features with images can substantially enhance overall performance as shown by Guillaumin et al. (2010) and Gong et al.(2014). While there are many research studies on audio-visual fusion for emotion recognition, just a few research works are dedicated to multimodal emotion or sentiment analysis using textual clues along with visual and audio modalities. Wollmer et al. and Rozgic et al. fused facts from audio, visual and textual modalities to extract emotion and sentiment. Metallinou et al. and Eyben et al. fused audio and textual modalities for emotion recognition. All the approaches depend on feature-level fusion. Wu et al. used decision level fusion for audio and textual clues. Nan Xu and Wenji Mao proposed a unified neural network model that utilizes CNN's to extract text and images clues .Note however that most of the previous work on audio-visual emotion analysis has focused exclusively on blending the audio and video modalities, and did not combine textual features, as we tend to liquidate our work.

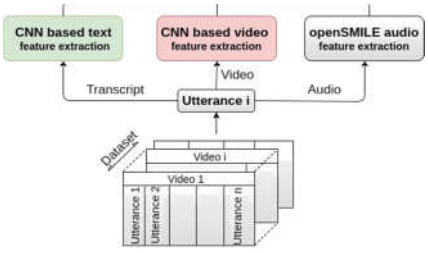
This analysis work suggests CNN-based structure for feature extraction of audio visual, text modality, and fusing them for tri-modal sentiment analysis and emotional recognition. This model exceeds the state of the art. Additionally, we tend to study the behavior of our technique within the aspects seldom addressed by different authors, like speaker independence, generalizability of the models and also the performance of individual modalities.

V. Method

I.Textual Features

The most unremarkably classification techniques like SVM, maximum Entropy and Naïve Bays are supported bag of words model within which the sequence of words is unnoticed. This may result in inefficient in mining the sentiment from the input with sequence of words. It may disturb the sentiment existing in it. By overcomingthis downside, several researches reportable by using deep learning in sentiment analysis. A deep neural network architecture that collectively uses wordlevel, character level and sentence level representationto perform sentiment analysis. For feature extraction from text data, we tend to use convolutional neural network (CNN). The trained CNN options were then fed into associate SVM for classification i.e., we used CNN as trainable feature extractor and SVM as a classifier.InFigure2, the concept behind convolution is to require the scalar product of vector of k weights wk, called kernel vector, with every k-gram within the sentence s(t) to get another sequence of options

c(t) = (c1(t); c2(t); : : : ; cL(t)):
cj = $wk^T xi:i+k$ --- (1)
we have a tendency to apply a max pooling operation over the feature map and take the maximum value
c^(t)=maxfc(t)g because
the feature adore this specific kernel vector and window sizes to get multiple features. For each word xi(t) within the vocabulary, A d-dimensional vector illustration, referred to as word embedding, was given during a look-up table that had been learned from the data. The vector illustration of a sentence is a concatenation of the vectors of individual words.



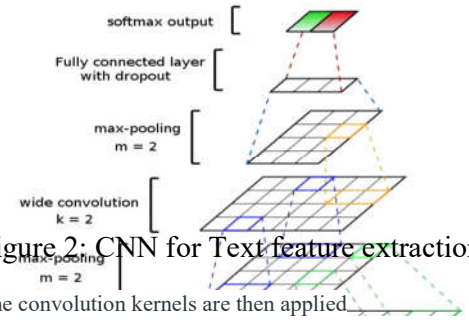


Figure 2: CNN for Text feature extraction

The convolution kernels are then applied to word vectors rather than individual words. Similarly, one will have look-up tables for features apart from words if these features are deemed useful. we have used these features to train higher layers of the CNN to represent larger teams of words in sentences. we denote the feature learned at a hidden neuron h in layer l as Fh1. Multiple features are learned in parallel at the identical CNN layer. The features learned at every layer are accustomed train the subsequent layer:
$F^l = X^{nh}{}_{h=1wkh\,Fl}$ ---(2)
where * denotes convolution, wk is a weight kernel for hidden neuron h and the sum of hidden neurons are nh. The CNN sentence model preserves the order of words by adopting convolution kernels of bit by bit increasing sizes, that span associate increasing variety of words and ultimately the entire sentence. Each word in a sentence was represented using word embedding.
We employed the publicly offered word2vec vectors that were trained on hundred billion of words from Google News. The vectors were of dimension d = 300, trained by continual bag-of-words frame work . Words not present in the set of pre-trained words were initialized arbitrarily. Each sentence was wrapped to a window of fifty words. Our CNN had 2 convolution layers. A kernel size of three and four, each one utilized fifty feature maps in the 1st convolution layer and a kernel size of two and hundred feature maps within the second layer . The convolution layers were interleaved with pooling layers of two dimensions. The activation value of 500-dimensional fully-connected layers of the network is our feature vector in the fusion process.

II.Audio-based Sentiment Analysis

The study of the relationship between emotional content and audio signals is a very mature field. Researchers have expanded the success found in using the MelFrequency Cepstral Coefficients (MFCC) for speech recognition community to explore their uses in music modeling. MFCCs are currently a staple in audio processing and are commonly used in MIR applications such as generic classification. Audio features like pitch, intensity and loudness are extracted using Open-EAR package and SVM classifier is used to detect the sentiment. The audio features are automatically extracted from the audio track of each video clip using OpenEAR package and Hidden Markov Models (HMM) classifier is made to detect the sentiment. Instead of extracting all the features from the entire input using tools like OpenEAR/OpenSMILE, only specific relevant features like MFCC, prosody and relative prosody are extracted from stressed and traditional regions of an input that used in our study. To observe the sentiment from natural audio streams, Maximum Entropy modeling and Part of Speech tagging are used to develop a sentiment detection system. Transcripts from audio streams are obtained using ASR(Automatic Speech Recognition). This approach shows that it is highly possible to automatically detect sentiment in natural spontaneous audio with good accuracy. Audio sentiment is generally extracted from the characteristics of the vocal tract, excitation and prosody.

III.Visual sentiment analysis

The fundamental evaluation responsibilities in “visible sentiment evaluation revolve round modeling, detective paintings and leveraging sentiment expressed by facial or physical gestures or sentiment related visual multimedia device. Due to the fact, the video facts could be very big accordingly creating our new frames for the video. For notably lessen the amount of education video information. The frames have been then exceeded via a CNN structure just like determine a video comprised of a sequence of pics. To capture the temporal dependence, we transformed each pair of consecutive pix at t and t + 1 into one picture and provided this transformed image as out input to the multi-level CNN. Thus, Pre-processing concerned scaling all video frames to half the resolution. Everycombine of consecutive video frames were reborn into one frame to attain temporal convolution options. All the frames were standardized to 250 500 pixels by padding with zeros. The first convolution layer contained one hundred kernels of size 10 20; the second one convolution layer had 100 kernels of size20 30; this layer became followed via a logistic layer of fully related 300 neurons and a softmax layer. The convolution layers had been included with pooling layers of dimension 2 2. The activation of the neurons inside the logistic layer become taken as the video features for the classification process

VI. Fusion

Multimodal Sentiment Analysis refers to the combination of two or more input modes to improvise the performance of the analysis. Combination of text and audio-visual inputs is an example of multimodal sentiment analysis.

Figure 3: The architecture of the proposed system

The fusion can be done in three ways Data fusion, Featurefusion and Decision fusion.Decision fusion is used in most of the work. Figure-3 is the whole structure of the proposed system. In feature level fusion, Thejoint feature vector is createdby the combination of multiple input features and send them as a combined vector to SVM for final decision then we discuss the result of this fusion in next session.

VII. Experiments and Observations

I.Datasets

Humans are sharing their opinions and reviews through online video sharing web sites a day. Studying sentiment and subjectivity in these opinion movies is experiencing a growing interest from academia and industry. Sentiment evaluation is successful for textual content.It is far an understudied studies place for multimedia content.

The biggest setbacks for learning in this path are lack of a right dataset, method, baselines and statistical evaluation of how information from different modality sources relate to each other.

II.MOSI

In this paper we introduce the first opinion-stage annotated corpus of sentiment and subjectivity evaluation in on-line videos called Multimodal Opinion-Stage Sentiment Intensity dataset (MOSI). It is constructed by Zadeh et al, consisting of 2199 video reviews of movies, books, and product. The dataset is rigorously annotated with labels for subjectivity, sentiment intensity, consistent with-frame and in keeping with-opinion annotated visible features, and in keeping with-milliseconds annotated audio features.

III.MOUD

For our experiments, we use the Multimodal Sentiment Analysis Datasets - MOUD developed by Perez-Rosas et al. They collected 100 product review and recommendation videos from YouTube. Each video became segmented into its utterances and each utterance becomes categorized by means of a sentiment (positive, negative and neutral). On average, each video has 6 utterances; every utterance is five seconds lengthy. Thedataset incorporates 498 utterances categorized tremendous, poor or neutral. In our test,we dropped the neutral label to maintain consistency with previous work.In this experiment to address thegeneralizability issues, the model is trained on MOSI and tested on MOUD. In this approach, the features collected from all the multimodal streams are combined into a single feature vector, thus resulting in one vector for each utterance in the dataset which is used to make a decision about the sentiment orientation of the utterance. Table-1 shows the performance of several comparative experiments, using one, two, and three modalities at a time. We use the entire set of 412 utterances and run 10 fold cross validations using an SVM classifier, as implemented in the Weka toolkit.

Table 1: Speaker dependent: cross validation result

Modality	MOSI	MOUD
Unimodal		
Text	75.16	48.40
Video	53.46	47.68
Audio	58.70	53.70
Bimodal		
Text+Audio	75.72	57.10
Text+Video	75.06	49.22
Video+Audio	62.40	62.88
Multimodal		
Text+Audio+Video	76.66	67.90

IV. Speaker-Independent Experiment

Maximum of the research in multimodal sentiment evaluation is accomplished on a datasets with speaker overlap in train and test splits. As we know, each individual has unique way of expressing his/her emotions and sentiment. Person independent sentiment features for analysis is highly important for all modalities. In real world applications, the model should be robust to person variance. Consequently, we carried out person-impartial experiments to emulate unseen situations. This time, our train/test splits of the datasets have been absolutely disjoint with appreciate to speakers.While checking out, our models had to classify feelings and sentiments from utterances by using speakers they have never visible before. Under, we enlist the process of this speaker-independent experiment using online video review data.

V. On line video review data

We commenced by collecting a hard and fast of movies from the social media web site YouTube, using numerous keywords likely to lead to a product evaluate or advice among all the videos back by way of the YouTube search, we selected best videos that reputable the subsequent suggestions: the speaker should be in the front of the camera; her face should be absolutely seen, with a minimum quantity of face occlusion at some point of the recording; there should not be any historical past tune or animation. The very final video set consists of a hundred videos randomly selected from the movies reviewfrom YouTube that still met the recommendations above. The dataset includes 100 speakers of 35 female and 65 male speakers, with their age approximately ranging from 20 to 60 years. All the videos were first pre-processed toeliminate introductory titles and advertisements. Since the reviewers often switched topics when expressing theiropinions, we manually selected a 30 seconds opinion segment from each video to avoid having multiple topics in a single review. As this dataset contains 100 speakers, we performed a 10 fold speaker independent test, where in each round one of the speaker was in the test set. The same SVM model was used and macro F score was used as a

metric forperformance improvement.Table-2 shows the result of speaker independent experiment. It can be seen that audio modality continuously performs higher than visual modality in all the dataset. The text mode plays the most essential role in both emotion recognition and sentiment

Table 2: Speaker independent: cross validation result

Modality	Macro F_Score
baseline	55.93
Unimodal	
Text	50.89
Video	41.60
Audio	45.60
Bimodal	
Text+Audio	51.70
Text+video	52.12
Video+Audio	46.35
Multimodal	
Text+Audio+video	52.44

analysis. The fusion of the modalities show more impact for emotional recognition than on sentiment analysis. As expected in all forms of experiments, bimodal and tri-modal have per-formed better than uni-modals, Overall, the performance of audio modality is better than visual on all the datasets. The unimodal performance of text modality is notably higher than other two modalities; Table 2 also presents the comparison with state of the art.

VI. Contributions of the Modalities

In order to understand the roles of eachmodality for overall classification,We have manually done some qualitative analysis of performance of all the

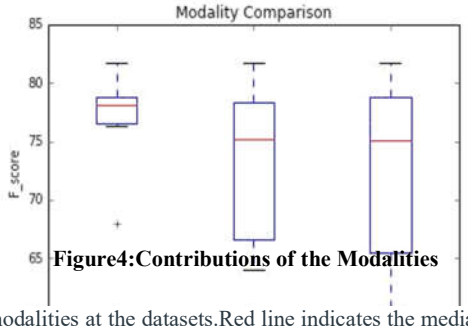


Figure4:Contributions of the Modalities

modalities at the datasets.Red line indicates the median F score value. Text classifications give the preference of high polar words, correctly identified the polarity as positive and helped the bi and multi-model features for correct classification. Text modality also helped in situation where face of the reviewers was not prominent. However in some utterance, text modality misclassified due to the presence of misleading linguistic cues. The presence of positive parses such as likes to see, responsible. However the high pitch of resentment in the person voice and glowering face aids to classify this to be a negative utterance.

VIII. Conclusion

The problem of detecting hidden emotion like sarcasm or irony has always challenged the researchers in this field. Since, these emotions are not directly expressed in the text. Combining multi-modal inputs with the focus on detecting hidden emotion is a future direction of research. In forthcoming work we plan to explore alternative multi-modal fusion method, such as decision-level and Meta level fusion to improve the integration of various modalities. Meanwhile, the main advantage of using CNN and support vector model is that we can transfer the procedure to other domains using a much simpler fine-tuning. Our future work shall explore other fusion strategies to further progress the performance of multimodal feature fusion, and test our models in the environment of more scalable social media big data.

References

[1]. Perez-Rosas, V., Mihalcea, R., Morency, Utterance-level multimodal senti-ment analysis. In: ACL (1). (2013)

[2 Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., Morency, L.P.: Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Systems 28 (2013)

[3]. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual fea-tures and multiple kernel learning for utterance-level multimodal sentiment anal-ysis. In: Proceedings of EMNLP. (2015)

[4]. M. Wang, D. Cao, L. Li, S. Li, and R. Ji, "Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model.," ICIMCS, pp. 76–80, 2014.

[5]. Cao D, Ji R, Lin D, et al. A cross-media public sentiment analysis system for microblog[J]. Multimedia Systems, 2014: 1-8.

[6]. Fuhai Chen, Yue Gao, Donglin Cao, and Rongrong Ji, "Multimodal hypergraph learning for microblog sentiment prediction," presented at the 2015 IEEE International Conference on Multimedia and Expo (ICME), 2015, pp. 1–6.

[7]. S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," Neurocomputing, vol. 174, pp. 50–59, Jan. 2016.

[8]. M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 2837–2841.

[9]. Basant Agarwal, SoujanyaPoria, Namita Mittal, Alexander Gelbukh, and Amir Hussain. 2015. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognitive Computation 7, 4 (2015), 487–499.

[10]. TadasBaltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In Applications of Computer Vision

(WACV), 2016 IEEE Winter Conference on. IEEE, 1–10.

[11]. Metallinou, A., Lee, S., Narayanan, S.: Audio-visual emotion recognition using gaussian mixture models for face and voice. In: Tenth IEEE International Sympo-sium on ISM 2008, IEEE (2008)

[12]. Eyben, F., W•ollmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R.: On-line emotion recognition in a 3-d activation-valence-time continuum us-ing acoustic and linguistic cues. Journal on Multimodal User Interfaces 3 (2010)

[13]. Wu, C.H., Liang, W.B.: Emotion recognition of aective speech based on multiple classiers using acoustic-prosodic information and semantic labels. IEEE Transac-tions on A ective Computing 2 (2011)

[14]. MatthieuGuillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In CVPR, 2010

[15]. DhirajJoshi, RitendraDatta, Elena Fedorovskaya, Quang-Tuan Luong, J.Z. Wang, Jia Li, and JieboLuo. Aesthetics and emotions in images. 28:94–115, 2011

[16]. Damian Borth, Tao Chen, Rong-Rong Ji, and Shih-Fu Chang. Sentibank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In ACM, 2013

[17] .Robert Plutchik. The nature of emotions. 89:344, 2001.

[18]. L. Kaushik, A. Sangwan, and J. H. Hansen. Automatic audio sentiment extraction using keyword spotting. In *Proc. INTERSPEECH*, pages 2709–2713, September 2015.

[19]. L. Kaushik, A. Sangwan, and J. H. L. Hansen. Sentiment extraction from natural audio streams.In *Proc.ICASSP*, pages 8485–8489, 2013.