Benchmarking Multimodal Sentiment Analysis

Erik Cambria[a], Devamanyu Hazarika[b], Soujanya Poria[c],
Amir Hussain[d], R.B.V. Subramanyam[b]

[a] School of Computer Science and Engineering, NTU, Singapore
[b] National Institute of Technology, Warangal, India
[c] Temasek Laboratories, NTU, Singapore
[d] School of Natural Sciences, University of Stirling, UK

Abstract. We propose a deep-learning-based framework for multimodal sentiment analysis and emotion recognition. In particular, we leverage on the power of convolutional neural networks to obtain a performance improvement of 10% over the state of the art by combining visual, text and audio features. We also discuss some major issues frequently ignored in multimodal sentiment analysis research, e.g., role of speaker-independent models, importance of different modalities, and generalizability. The framework illustrates the different facets of analysis to be considered while performing multimodal sentiment analysis and, hence, serves as a new benchmark for future research in this emerging field.

Key words: Multimodal Sentiment Analysis, Emotion Detection, Deep Learning, Convolutional Neural Networks.

1 Introduction

Emotion recognition and sentiment analysis have become a new trend in social media analytics because of the immense opportunities they offer in terms of understanding preferences and habits of users and their contents [1]. With the advancement of communication technology, abundance of smartphones and the rapid rise of social media, a larger and larger amount of data is being uploaded in video, rather than text, format [2]. For example, consumers tend to record their reviews and opinions on products using a web camera and upload them on social media platforms such as YouTube or Facebook to inform subscribers of their views. Such videos often contain comparisons of products from competing brands, pros and cons of product specifications, and other information that can aid prospective buyers to make informed decisions.

The primary advantage of analyzing videos over mere text analysis for detecting emotions and sentiment from opinions is the surplus of behavioral cues. Video provides multimodal data in terms of vocal and visual modalities. The vocal modulations and facial expressions in the visual data, along with text data, provide important cues to better identify true affective states of the opinion holder. Thus, a combination of text and video data helps to create a better emotion and sentiment analysis model.

Recently, a number of approaches to multimodal sentiment analysis producing interesting results have been proposed [3–7]. However, there are major issues that remain unaddressed in this field, such as the role of speaker-dependent and speaker-independent models, the impact of each modality across datasets, and generalization ability of a multimodal sentiment classifier. Not tackling these issues has presented difficulties in effective comparison of different multimodal sentiment analysis methods. In this paper, we address some of these issues and, in particular, propose a novel framework that outperforms the state of the art on benchmark datasets by more than 10%. We use a deep convolutional neural network (CNN) to extract features from visual and text modalities.

The paper is organized as follows: Section 2 provides a brief literature review on multimodal sentiment analysis; Section 3 presents the proposed framework; experimental results and discussion are given in Section 4; Section 5 proposes a qualitative analysis; finally, Section 6 concludes the paper.

2 Related Work

Text-based sentiment analysis systems can be broadly categorized into knowledge-based and statistics-based systems [8]. While the use of knowledge bases was initially more popular for the identification of emotions and polarity in text [9, 10], sentiment analysis researchers have recently been using statistics-based approaches, with a special focus on supervised statistical methods [11–13].

In 1970, Ekman et al. [14] carried out extensive studies on facial expressions. Their research showed that universal facial expressions are able to provide sufficient clues to detect emotions. Recent studies on speech-based emotion analysis [15] have focused on identifying relevant acoustic features, such as fundamental frequency (pitch), intensity of utterance, bandwidth, and duration.

As to fusing audio and visual modalities for emotion recognition, two of the early works were done by De Silva et al. [16] and Chen et al. [17]. Both works showed that a bimodal system yielded a higher accuracy than any unimodal system. More recent research on audio-visual fusion for emotion recognition has been conducted at either feature level [18] or decision level [19].

While there are many research papers on audio-visual fusion for emotion recognition, only a few research works have been devoted to multimodal emotion or sentiment analysis using text clues along with visual and audio modalities. Wollmer et al. [4] and Rozgic et al. [20] fused information from audio, visual and text modalities to extract emotion and sentiment. Metallinou et al. [21] and Eyben et al. [22] fused audio and text modalities for emotion recognition. Both approaches relied on feature-level fusion. Wu et al. [23] fused audio and textual clues at decision level.

In this paper, we propose CNN-based framework for feature extraction from visual and text modality and a method for fusing them for multimodal sentiment analysis. In addition, we study the behavior of our method in the aspects rarely addressed by other authors, such as speaker independence, generalizability of the models and performance of individual modalities.

3 Method

3.1 Textual Features

For feature extraction from textual data, we used a CNN. The trained CNN features were then fed into a support vector machine (SVM) for classification, i.e., we used CNN as trainable feature extractor and SVM as a classifier (Fig. 1).

The idea behind convolution is to take the dot product of a vector of k weights w_k, known as kernel vector, with each k-gram in the sentence $s(t)$ to obtain another sequence of features $c(t) = (c_1(t), c_2(t), \ldots, c_{\mathrm{L}}(t))$:

$$c_j = w_k^T \cdot \mathbf{x}_{i:i+k-1}. \tag{1}$$

We then apply a max pooling operation over the feature map and take the maximum value $\hat{c}(t) = \max\{\mathbf{c}(t)\}$ as the feature corresponding to this particular kernel vector. We used varying kernel vectors and window sizes to obtain multiple features.

For each word $x_i(t)$ in the vocabulary, a d-dimensional vector representation, called word embedding, was given in a look-up table that had been learned from the data [24]. The vector representation of a sentence was a concatenation of the vectors for individual words. The convolution kernels are then applied to word vectors instead of individual words. Similarly, one can have look-up tables for features other than words if these features are deemed helpful.

We used these features to train higher layers of the CNN to represent bigger groups of words in sentences. We denote the feature learned at a hidden neuron h in layer l as F_h^l. Multiple features are learned in parallel at the same CNN layer. The features learned at each layer are used to train the next layer:

$$F^l = \sum\nolimits_{h=1}^{n_h} w_k^h * F^{l-1}, \tag{2}$$

where * denotes convolution, w_k is a weight kernel for hidden neuron h and n_h is the total number of hidden neurons. The CNN sentence model preserves the order of words by adopting convolution kernels of gradually increasing sizes, which span an increasing number of words and ultimately the entire sentence.

Each word in a sentence was represented using word embeddings. We employed the publicly available word2vec vectors, which were trained on 100 billion words from Google News. The vectors were of dimensionality $d = 300$, trained using the continuous bag-of-words architecture [24]. Words not present in the set of pre-trained words were initialized randomly.

Each sentence was wrapped to a window of 50 words. Our CNN had two convolution layers. A kernel size of 3 and 4, each of them having 50 feature maps was used in the first convolution layer and a kernel size 2 and 100 feature maps in the second one. We used ReLU as the non-linear activation function of the network. The convolution layers were interleaved with pooling layers of dimension 2. We used the activation values of the 500-dimensional fully-connected layer of the network as our feature vector in the final fusion process.

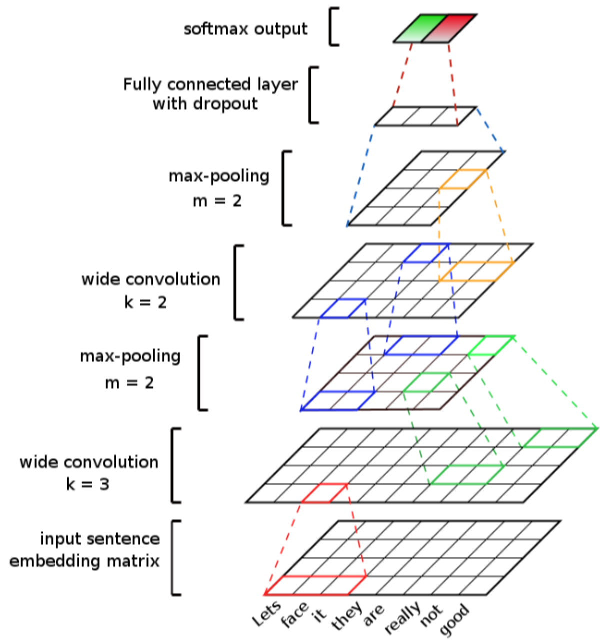

Fig. 1: CNN for feature extraction from text modality.

3.2 Audio Features

We automatically extracted audio features from each annotated segment of the videos. Audio features were also extracted in 30 Hz frame-rate; we used a sliding window of 100 ms. To compute the features, we used the open-source software openSMILE [25]. This toolkit automatically extracts pitch and voice intensity. Voice normalization was performed and voice intensity was thresholded to identify samples with and without voice. Z-standardization was used to perform voice normalization.

The features extracted by openSMILE consist of several low-level descriptors (LLD) and their statistical functionals. Some of the functionals are amplitude mean, arithmetic mean, root quadratic mean, etc. Taking into account all functionals of each LLD, we obtained 6373 features.

3.3 Visual Features

Since the video data is very large, we only considered every tenth frame in our training videos. The constrained local model (CLM) was used to find the outline of the face in each frame [26]. The cropped frame size was further reduced by scaling down to a lower resolution, thus creating our new frames for the video. In this way, we could drastically reduce the amount of training video data. The frames were then passed through a CNN architecture similar to Fig. 1.

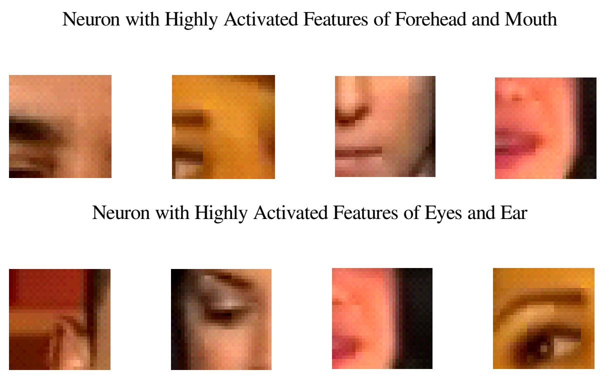

Fig. 2: Top image segments activated at two feature detectors in the first layer of a deep CNN

To capture the temporal dependence of the images constituting the video, we transformed each pair of consecutive images at t and $t+1$ into a single image and provided this transformed image as input to the multilevel CNN. We used kernels of varying dimensions to learn Layer-1 2D features (shown in Fig. 2) from the transformed input. Similarly, the second layer also used kernels of varying dimensions to learn 2D features. The down-sampling layer transformed features of different kernel sizes into uniform 2D features and was then followed by a logistic layer of neurons.

Pre-processing involved scaling all video frames to half of their resolution. Each pair of consecutive video frames were converted into a single frame to achieve temporal convolution features. All frames were standardized to 250×500 pixels by padding with zeros.

The first convolution layer contained 100 kernels of size 10×20; the next convolution layer had 100 kernels of size 20×30; this layer was followed by a logistic layer of fully connected 300 neurons and a softmax layer. The convolution layers were interleaved with pooling layers of dimension 2×2. The activation of the neurons in the logistic layer were taken as the video features for the classification task.

3.4 Fusion

In order to fuse the information extracted from each modality, we concatenated feature vectors extracted from each modality and sent the combined vector to a SVM for the final decision. This scheme of fusion is called feature-level fusion. Since the fusion involved concatenation and no overlapping merge or combination, scaling and normalization of the features were avoided. We discuss the results of this fusion in Section 4. The overall architecture of the proposed method can be seen in Fig. 3.

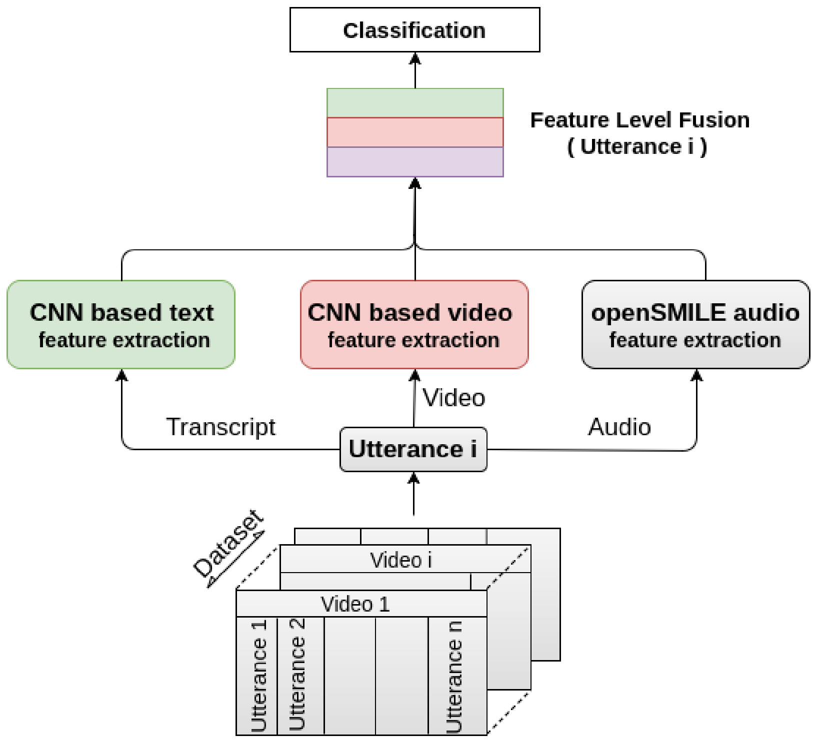

Fig. 3: Overall architecture of the proposed method.

4 Experiments and Observations

4.1 Datasets

Multimodal Sentiment Analysis Datasets For our experiments, we used the MOUD dataset, developed by Perez-Rosas et al. [3]. They collected 80 product review and recommendation videos from YouTube. Each video was segmented into its utterances (498 in total) and each of these was categorized by a sentiment label (positive, negative and neutral). On average, each video has 6 utterances and each utterance is 5 seconds long. In our experiment, we did not consider neutral labels, which led to the final dataset consisting of 448 utterances. We dropped the neutral label to maintain consistency with previous work. In a similar fashion, Zadeh et al. [27] constructed a multimodal sentiment analysis dataset called multimodal opinion-level sentiment intensity (MOSI), which is bigger than MOUD, consisting of 2199 opinionated utterances, 93 videos by 89 speakers. The videos address a large array of topics, such as movies, books, and products. In the experiment to address the generalizability issues, we trained a model on MOSI and tested on MOUD.

Multimodal Emotion Recognition Dataset The IEMOCAP database [28] was collected for the purpose of studying multimodal expressive dyadic interactions. This dataset contains 12 hours of video data split into 5 minutes of dyadic interaction between professional male and female actors. Each interaction session was split into spoken utterances. At least 3 annotators assigned to each utterance one emotion category: *happy, sad, neutral, angry, surprised, excited, frustration, disgust, fear* and *other*. In this work, we considered only the utterances with majority agreement (i.e., at least two out of three annotators labeled the same emotion) in the emotion classes of *angry*, *happy*, *sad*, and *neutral*. We take only these four classes for comparison with the state of the art [29] and other authors.

4.2 Speaker-Independent Experiment

Most of the research in multimodal sentiment analysis is performed on datasets with speaker overlap in train and test splits. Given this overlap, however, results do not scale to true generalization. In real-world applications, the model should be robust to person variance. Thus, we performed person-independent experiments to emulate unseen conditions. This time, our train/test splits of the datasets were completely disjoint with respect to speakers. While testing, our models had to classify emotions and sentiments from utterances by speakers they have never seen before. Below, we enlist the procedure of this speaker-independent experiment:

- **IEMOCAP:** As this dataset contains 10 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 as before and macro F-score was used as a metric.

- **MOUD:** This dataset contains videos of about 80 people reviewing various products in Spanish. Each utterance in the video has been labeled as *positive, negative* or *neutral*. In our experiments, we consider only *positive* and *negative* sentiment labels. The speakers were divided into 5 groups and a 5-fold person-independent experiment was run, where in every fold one out of the five group was in the test set. Finally, we took average of the macro F-score to summarize the results (Table 1).

- **MOSI:** The MOSI dataset is a dataset rich in sentimental expressions where 93 people review topics in English. The videos are segmented with each segment's sentiment label scored between $+3$ to -3 by 5 annotators. We took the average of these labels as the sentiment polarity and, hence, considered only two classes (*positive* and *negative*). Like MOUD, speakers were divided into 5 groups and a 5-fold person-independent experiment was run. During each fold, about 75 people were in the training set and the remaining in the test set. The training set was further split randomly into 80%–20% and shuffled to generate train and validation splits for parameter tuning.

Comparison with the Speaker-Dependent Experiment In comparison with the speaker-dependent experiment, the speaker-independent experiment performs poorly. This is due to the lack of knowledge about speakers in the dataset. Table 2 shows the performance obtained in the speaker-dependent experiment. It can be seen that audio modality consistently performs better than visual modality in both MOSI and IEMOCAP datasets. The text modality plays the most important role in both emotion recognition and sentiment analysis. The fusion of the modalities shows more impact for emotion recognition than for sentiment analysis. Root mean square error (RMSE) and TP-rate of the experiments using different modalities on IEMOCAP and MOSI datasets are shown in Fig. 4.

Modality	Source	IEMOCAP	MOUD	MOSI
Unimodal	A	51.52	53.70	57.14
	V	41.79	47.68	58.46
	T	65.13	48.40	75.16
Bimodal	T + A	70.79	57.10	75.72
	T + V	68.55	49.22	75.06
	A + V	52.15	62.88	62.4
Multimodal	T + A + V	**71.59**	**67.90**	**76.66**

Table 1: **Speaker-Independent:** Macro F-score reported for speaker-independent classification. *IEMOCAP:* 10-fold speaker-independent average. *MOUD:* 5-fold speaker-independent average. *MOSI:* 5-fold speaker-independent average. *Legenda:* A stands for Audio, V for Video, T for Text.

Modality	Source	IEMOCAP	MOSI
Unimodal	Audio	66.20	64.00
	Video	60.30	62.11
	Text	67.90	78.00
Bimodal	Text + Audio	78.20	76.60
	Text + Video	76.30	78.80
	Audio + Video	73.90	66.65
Multimodal	Text + Audio + Video	**81.70**	**78.80**
	Text + Audio + Video	**69.35**[1]	**73.55**[2]

Table 2: **Speaker-Dependent:** Ten-fold cross-validation results on IEMOCAP dataset and 5-fold CV results (macro F-score) on MOSI dataset. [1]By [29]; [2]by [5].

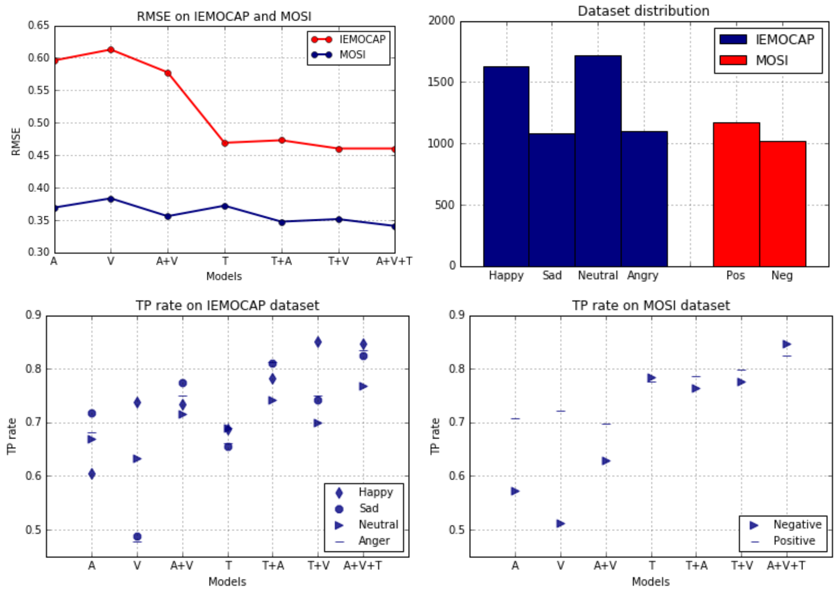

Fig. 4: Experiments on IEMOCAP and MOSI datasets. The top-left figure shows the RMSE of the models on IEMOCAP and MOSI. The top-right figure shows the dataset distribution. Bottom-left and bottom-right figures present TP-rate on of the models on IEMOCAP and MOSI dataset, respectively.

4.3 Contributions of the Modalities

As expected, bimodal and trimodal models have performed better than unimodal models in all experiments. Overall, audio modality has performed better than visual on all datasets. Except for MOUD dataset, the unimodal performance of text modality is notably better than other two modalities (Fig. 5). Table 2 also presents the comparison with state of the art. The present method outperformed the state of the art by 12% and 5% on the IEMOCAP and MOSI datasets, respectively.[1] The method proposed by Poria et al. is similar to ours, except for the fact they used a standard CLM-based facial feature extraction method. Hence, our proposed CNN-based visual feature extraction algorithm has helped to outperform the method by Poria et al.

4.4 Generalizability of the Models

To test the generalization ability of the models, we have trained the framework on MOSI dataset in speaker-independent fashion and tested on MOUD dataset. From Table 3, we can see that the trained model on MOSI dataset performed poorly on MOUD dataset.

[1] We have reimplemented the method by Poria et al. [5]

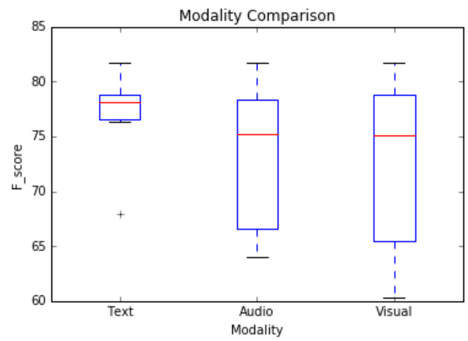

Fig. 5: Performance of the modalities on the datasets. Red line indicates the median of the F-score.

Modality	Source	Macro F-score
Unimodal	Audio	41.60 %
	Video	45.50 %
	Text	50.89 %
Bimodal	Text + Audio	51.70 %
	Text + Video	52.12 %
	Audio + Video	46.35 %
Multimodal	Text + Audio + Video	52.44 %

Table 3: **Cross-dataset results:** Model (with previous configurations) trained on MOSI dataset and tested on MOUD dataset.

This is mainly due to the fact that reviews in MOUD dataset had been recorded in Spanish so both audio and text modalities miserably fail in recognition, as MOSI dataset contains reviews in English. A more comprehensive study would be to perform generalizability tests on datasets in the same language. However, we were unable to do this for the lack of benchmark datasets. Also, similar experiments of cross-dataset generalization was not performed on emotion detection given the availability of only a single dataset (IEMOCAP).

4.5 Visualization of the Datasets

MOSI visualizations present information regarding dataset distribution within single and multiple modalities (Fig. 6). For the textual and audio modalities, comprehensive clustering can be seen with substantial overlap. However, this problem is reduced in the video and all modalities with structured declustering but overlap is reduced only in multimodal. This forms an intuitive explanation of the improved performance in the multimodality. IEMOCAP visualizations provide insight for the 4-class distribution for uni and multimodals, where clearly the multimodal distribution has the least overlap (increase in red and blue visuals, apart from the rest) with sparse distribution aiding the classification process.

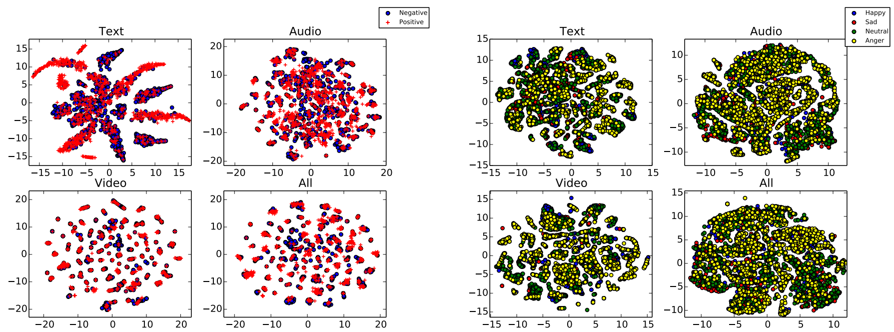

Fig. 6: T-SNE 2D visualization of MOSI and IEMOCAP datasets when unimodal features and multimodal features are used.

5 Qualitative Analysis

In order to have a better understanding of roles of modalities for the overall classification, we performed a qualitative analysis. Here, we show the cases where our model successfully comprehends the semantics of the utterances and, with aid from the multiple media, correctly classifies the emotion of the same.

While overviewing the correctly classified utterances in the validation set, we found out that text modality often helped the classification of utterances where visual and audio cues were flat with less variance. In such situations, the model gathered information from the language semantics extracted by the text modality. For example, in an utterance from the MOSI dataset "amazing special effects", there was no jest of enthusiasm in speaker's voice and face audio-visual classifier, which caused failure to identify the positivity of this utterance by the audio and video unimodal classifiers. The text classifier, instead, correctly detected the polarity as positive (given the presence of highly polar words) and, hence, helped the bimodal and multimodal classifiers to correctly classify the utterance.

The text modality also helped in situations where the face of the reviewer was not visible (which happens quite often in product reviews). Even in cases where the text modality led to a misclassification (e.g., due to the presence of misleading linguistic cues), the overall classification was correct thanks to the audio and video inputs. For example, the text classifier classified the sentence "that like to see comic book characters treated responsibly" as positive (possibly because of the presence of positive phrases such as "like to see" and "responsibly"); however, the high pitch of anger in the person's voice and the frowning face helps to identify this as a negative utterance.

The above examples demonstrate the effectiveness and robustness of our model to capture overall video semantics of the utterances for emotion and sentiment detection. They also show how bimodal and multimodal models overcome the limitations of unimodal networks, given the multiple media input.

We also explored the misclassified validation utterances and found some interesting trends. Most videos consist of a group of utterances that have contextual dependencies among them. Thus, our model failed to classify utterances whose emotional polarity was highly dependent on the context described in an earlier or later part of the video. The modeling of such an inter-dependence, however, was out of the scope of this paper and, hence, we left it to future work.

6 Conclusion

We have presented a framework (available as demo[2]) for multimodal sentiment analysis and multimodal emotion recognition, which outperforms the state of the art in both tasks by a significant margin. We also discussed some major aspects of multimodal sentiment analysis problem such as the performance of speaker-independent models and cross-dataset performance of the models.

Our future work will focus on extracting semantics from the visual features, relatedness of the cross-modal features and their fusion. We will also include contextual dependency learning in our model to overcome the limitations mentioned in the previous section.

References

1. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37** (2017) 98–125
2. Cambria, E., Wang, H., White, B.: Guest editorial: Big social data analysis. Knowledge-Based Systems **69** (2014) 1–2
3. Pérez-Rosas, V., Mihalcea, R., Morency, L.P.: Utterance-level multimodal sentiment analysis. In: ACL. (2013) 973–982
4. 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) 46–53
5. Poria, S., Cambria, E., Gelbukh, A.: Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proceedings of EMNLP. (2015) 2539–2544
6. Zadeh, A.: Micro-opinion sentiment intensity analysis and summarization in online videos. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM (2015) 587–591
7. Poria, S., Chaturvedi, I., Cambria, E., Hussain, A.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: ICDM, Barcelona (2016) 439–448
8. Cambria, E.: Affective computing and sentiment analysis. IEEE Intelligent Systems **31** (2016) 102–107
9. Cambria, E., Poria, S., Bajpai, R., Schuller, B.: SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In: COLING. (2016) 2666–2677

[2] http://sentic.net/demo (best viewed in Mozilla Firefox)

10. Poria, S., Gelbukh, A., Agarwal, B., Cambria, E., Howard, N. Advances in Soft Computing and Its Applications. In: Common sense knowledge based personality recognition from text. Springer (2013) 484–496
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of ACL. (2002) 79–86
12. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of EMNLP. Volume 1631. (2013) 1642
13. Oneto, L., Bisio, F., Cambria, E., Anguita, D.: Statistical learning theory and ELM for big social data analysis. IEEE Computational Intelligence Magazine **11** (2016) 45–55
14. Ekman, P.: Universal facial expressions of emotion. Culture and Personality: Contemporary Readings/Chicago (1974)
15. Datcu, D., Rothkrantz, L.: Semantic audio-visual data fusion for automatic emotion recognition. Euromedia (2008)
16. De Silva, L.C., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multimodal information. In: Proceedings of ICICS. Volume 1., IEEE (1997) 397–401
17. Chen, L.S., Huang, T.S., Miyasato, T., Nakatsu, R.: Multimodal human emotion/expression recognition. In: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE (1998) 366–371
18. Kessous, L., Castellano, G., Caridakis, G.: Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces **3** (2010) 33–48
19. Schuller, B.: Recognizing affect from linguistic information in 3d continuous space. IEEE Transactions on Affective Computing **2** (2011) 192–205
20. Rozgic, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Prasad, R.: Ensemble of SVM trees for multimodal emotion recognition. In: IEEE APSIPA. (2012) 1–4
21. Metallinou, A., Lee, S., Narayanan, S.: Audio-visual emotion recognition using gaussian mixture models for face and voice. In: Tenth IEEE International Symposium on ISM 2008, IEEE (2008) 250–257
22. Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., Cowie, R.: On-line emotion recognition in a 3d activation-valence-time continuum using acoustic and linguistic cues. Journal on Multimodal User Interfaces **3** (2010) 7–19
23. Wu, C.H., Liang, W.B.: Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Transactions on Affective Computing **2** (2011) 10–21
24. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
25. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia, ACM (2010) 1459–1462
26. Baltrušaitis, T., Robinson, P., Morency, L.P.: 3D constrained local model for rigid and non-rigid facial tracking. In: IEEE CVPR. (2012) 2610–2617
27. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems **31** (2016) 82–88
28. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42** (2008) 335–359
29. Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Prasad, R.: Ensemble of svm trees for multimodal emotion recognition. In: IEEE APSIPA. (2012) 1–4