Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines

Soujanya Poria
Nanyang Technological University

Navonil Majumder
Instituto Politécnico Nacional

Devamanyu Hazarika
National University of Singapore

Erik Cambria
Nanyang Technological University

Alexander Gelbukh
Instituto Politécnico Nacional

Amir Hussain
Edinburgh Napier University

Editor:
Erik Cambria
cambria@ntu.edu.sg

We compile baselines, along with dataset split, for multimodal sentiment analysis. In this paper, we explore three different deep-learning-based architectures for multimodal sentiment classification, each improving upon the previous. Further, we evaluate these architectures with multiple datasets with fixed train/test partition. We also discuss some major issues, frequently ignored in multimodal sentiment analysis research, e.g., the role of speaker-exclusive models, the importance of different modalities, and generalizability. This 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.

Emotion recognition and sentiment analysis is opening up numerous opportunities pertaining to social media in terms of understanding users' preferences, habits, and their contents.[10] With the advancement of communication technology, an abundance of mobile devices, and the rapid rise of social media, a large amount of data is being uploaded as a video, rather than text.[2] For example, consumers tend to record their opinions on products using a webcam and upload them on social media platforms, such as YouTube and Facebook, to inform the 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.

Published by the IEEE Computer Society
1541-1672/18/$33.00 ©2018 IEEE

The primary advantage of analyzing videos over mere text analysis, for detecting emotions and sentiment, is the surplus of behavioral cues. Videos provide 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.[11,13] However, there are major issues that remain mostly unaddressed in this field, such as the consideration of the context in classification, effect of speaker-inclusive and speaker-exclusive scenario, the impact of each modality across datasets, and generalization ability of a multimodal sentiment classifier. Not tackling these issues has presented difficulties in the effective comparison of different multimodal sentiment analysis methods. In this paper, we outline some methods that address these issues and set up a baseline based on state-of-the-art methods. We use a deep convolutional neural network (CNN) to extract features from visual and text modalities.

This paper is organized as follows: The "Related Work" section provides a brief literature review on multimodal sentiment analysis. The "Unimodal Feature Extraction" section briefly discusses the baseline methods; experimental results and discussion are given in the "Experiments and Observations" section, and finally, "Conclusion" section concludes the paper.

RELATED WORK

In 1970, Ekman *et al*.[6] carried out extensive studies on facial expressions. Their research work showed that universal facial expressions are able to provide sufficient clues to detect emotions. Recent studies on speech-based emotion analysis[4] have focused on identifying relevant acoustic features, such as fundamental frequency (pitch), the 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*.[5] and Chen *et al*.[3] Both works showed that a bimodal system yielded a higher accuracy than any unimodal system.

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*.[14] fused information from audio, visual and text modalities to extract emotion and sentiment. Metallinou *et al.*[8] fused audio and text modalities for emotion recognition. Both approaches relied on feature-level fusion.

In this paper, we study the behavior of the method proposed in Poria *et al.*,[12] in the aspects rarely addressed by other authors, such as speaker independence, the generalizability of the models and performance of individual modalities.

UNIMODAL FEATURE EXTRACTION

For the unimodal feature extraction, we follow the procedures by bc-LSTM.[12]

Textual Feature Extraction

We employ CNN for textual feature extraction. Following Kim,[16] we obtain *n*-gram features from each utterance using three distinct convolution filters of sizes 3, 4, and 5, respectively, each having 50 feature-maps. Outputs are then subjected to max-pooling followed by rectified linear unit activation. These activations are concatenated and fed to a 100-dimensional (100-D) dense layer, which is regarded as the textual utterance representation. This network is trained at utterance level with the emotion labels.

Audio and Visual Feature Extraction

Identical to Poria *et al.*,[12] we use 3-D-CNN and openSMILE[7] for visual and acoustic feature extraction, respectively.

Fusion

In order to fuse the information extracted from different modalities, we concatenated the feature vectors representative of the given modalities and sent the combined vector to a classifier for the classification. 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 the "Experiments and Observations" section.

Baseline Method

1. *bc-LSTM:* We follow the method bc-LSTM[12] where they used a bidirectional LSTM to capture the context from the surrounding utterances to generate context-aware utterance representation.
2. *SVM:* After extracting the features, we merged and sent to an SVM with RBF kernel for the final classification.

EXPERIMENTS AND OBSERVATIONS

In this section, we discuss the datasets and the experimental settings. Also, we analyze the results yielded by the aforementioned methods.

Datasets

1. *Multimodal Sentiment Analysis Datasets:* For our experiments, we used the MOUD dataset, developed by Perez-Rosas *et al*.[9] 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 six utterances and each utterance is five 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*.[15] 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. Table 1 shows the split of train/test of these datasets.
2. *Multimodal Emotion Recognition Dataset:* The IEMOCAP database[1] was collected for the purpose of studying multimodal expressive dyadic interactions. This dataset contains 12 hours of video data split into five minutes of dyadic interaction between professional male and female actors. Each interaction session was split into spoken utterances. At least three annotators

Table 1. Person-independent train/test split details of each dataset (≈70/30% split). Note: $X \rightarrow Y$ represents train: X and test: Y; Validation sets are extracted from the shuffled train sets using 80/20% train/val ratio.

Dataset	Train		Test	
	utterance	***video***	***utterance***	***video***
IEMOCAP	4290	120	1208	31
MOSI	1447	62	752	31
MOUD	322	59	115	20
MOSI → MOUD	2199	93	437	79

assigned to each utterance one emotion category: *happy, sad, neutral, angry, surprised, excited, frustration, disgust, fear* and *other*. In this paper, 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*. Table 1 shows the split of train/test of this dataset.

Speaker-Exclusive Experiment

Most of the research work on multimodal sentiment analysis is performed with datasets having a common speaker(s) between train and test splits. However, given this overlap, results do not scale to true generalization. In real-world applications, the model should be robust to speaker variance. Thus, we performed speaker-exclusive 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 elaborate this speaker-exclusive experiment:

- **IEMOCAP:** As this dataset contains ten speakers, we performed a ten-fold speaker-exclusive test, where in each round exactly one of the speakers was included in the test set and missing from the train set. The same SVM model was used as before and accuracy was used as a performance 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 samples with *positive* and *negative* sentiment labels. The speakers were partitioned into five groups and a five-fold person-exclusive experiment was performed, where in every fold one out of the five group was in the test set. Finally, we took an average of the accuracy to summarize the results (Table 2).
- **MOSI:** MOSI dataset is rich in sentimental expressions, where 93 people review various products in English. The videos are segmented into clips, where each clip is assigned a sentiment score between 3 to +3 by five annotators. We took the average of these labels as the sentiment polarity and naturally considered two classes (*positive* and *negative*). Like MOUD, speakers were divided into five groups and a five-fold person-exclusive experiment was run. For each fold, on average 75 people were in the training set and the remaining in the test set. The training set was further partitioned and shuffled into 80%– 20% split to generate train and validation sets for parameter tuning.

Table 2. Accuracy reported for speaker-exclusive (Sp-Ex) and speaker-inclusive (Sp-In) split for Concatenation- Based Fusion. *IEMOCAP:* 10-fold speaker-exclusive average. *MOUD:* Five-fold speaker-exclusive average. *MOSI:* Five-fold speaker-exclusive average. *Legend:* A stands for Audio, V for Video, T for Text.

Modality	IEMOCAP		MOUD		MOSI	
Combi-nation	Sp-In	Sp-Ex	Sp-In	Sp-Ex	Sp-In	Sp-Ex
A	66.20	51.52	–	53.70	64.00	57.14
V	60.30	41.79	–	47.68	62.11	58.46
T	67.90	65.13	–	48.40	78.00	75.16
T + A	78.20	70.79	–	57.10	76.60	75.72
T + V	76.30	68.55	–	49.22	78.80	75.06
A + V	73.90	52.15	–	62.88	66.65	62.4
T + A + V	**81.70**	**71.59**	–	**67.90**	**78.80**	**76.66**

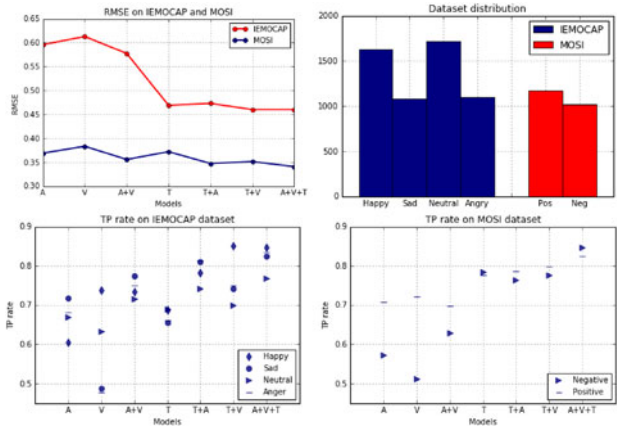

Figure 1. Experiments on IEMOCAP and MOSI datasets. The top-left figure shows the RMSE of the models on IEMOCAP and MOSI. The top-right figure shows the dataset distribution. Bottom-left and bottom-right figures present TP-rate of the models on IEMOCAP and MOSI dataset, respectively.

1) Speaker-Inclusive versus Speaker-Exclusive: In comparison with the speaker-inclusive experiment, the speaker-exclusive setting yielded inferior results. This is caused by the absence of knowledge about the speakers during the testing phase. Table 2 shows the performance obtained in the speaker-inclusive 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 Figure 1.

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 substantially better than the other two modalities (Figure 2).

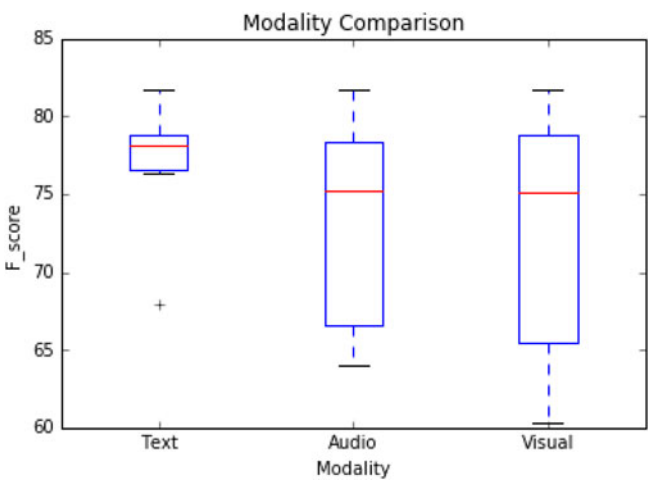

Figure 2. Performance of the modalities on the datasets. Red line indicates the median of the accuracy.

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

Modality Combination	Accuracy	
	SVM	bc-LSTM
T	46.5%	**46.9%**
V	43.3%	**49.6%**
A	42.9%	**47.2%**
T + A	50.4%	**51.3%**
T + V	49.8%	**49.8%**
A + V	46.0%	**49.6%**
T + A + V	51.1%	**52.7%**

Generalizability of the Models

To test the generalization ability of the models, we trained the framework on MOSI dataset in speaker-exclusive fashion and tested with MOUD dataset. From Table 3, we can see that the trained model with MOSI dataset performed poorly with 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 of the same language. However, we were unable to do this for the lack of benchmark datasets. Also, similar experiments of cross-dataset generalization were not performed on emotion detection, given the availability of only a single dataset (IEMOCAP).

Table 4. Accuracy reported for speaker-exclusive classification. *IEMOCAP:* Ten-fold speaker-exclusive average. *MOUD:* Five-fold speaker-exclusive average. *MOSI:* 5-fold speaker-exclusive average. *Legend:* A represents Audio, V represents Video, T represents Text.

Modality	IEMOCAP		MOUD		MOSI	
Combination	SVM	bc-LSTM	SVM	bc-LSTM	SVM	bc-LSTM
A	52.9	**57.1**	51.5	**59.9**	58.5	**60.3**
V	47.0	**53.2**	46.3	**48.5**	53.1	**55.8**
T	65.5	**73.6**	49.5	**52.1**	75.5	**78.1**
T + A	70.1	**75.4**	53.1	**60.4**	75.8	**80.2**
T + V	68.5	**75.6**	50.2	**52.2**	76.7	79.3
A + V	67.6	**68.9**	62.8	**65.3**	58.6	62.1
T + A + V	72.5	**76.1**	66.1	**68.1**	77.9	80.3

Comparison Among the Baseline Methods

Table 4 consolidates and compares the performance of all the baseline methods for all the datasets. We evaluated SVM and bc-LSTM fusion with MOSI, MOUD, and IEMOCAP dataset.

From Table 4, it is clear that bc-LSTM performs better than SVM across all the experiments. So, it is very apparent that consideration of the context in the classification process has substantially boosted the performance.

Visualization of the Datasets

MOSI visualizations present information regarding dataset distribution within single and multiple modalities (Figure 3). 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 scenario with structured declustering but the overlap is reduced only in multimodal. This forms an intuitive explanation of the improved performance in the multimodal scenario. IEMOCAP visualizations provide insight for the four-class distribution for uni- and multimodal scenario, where clearly the

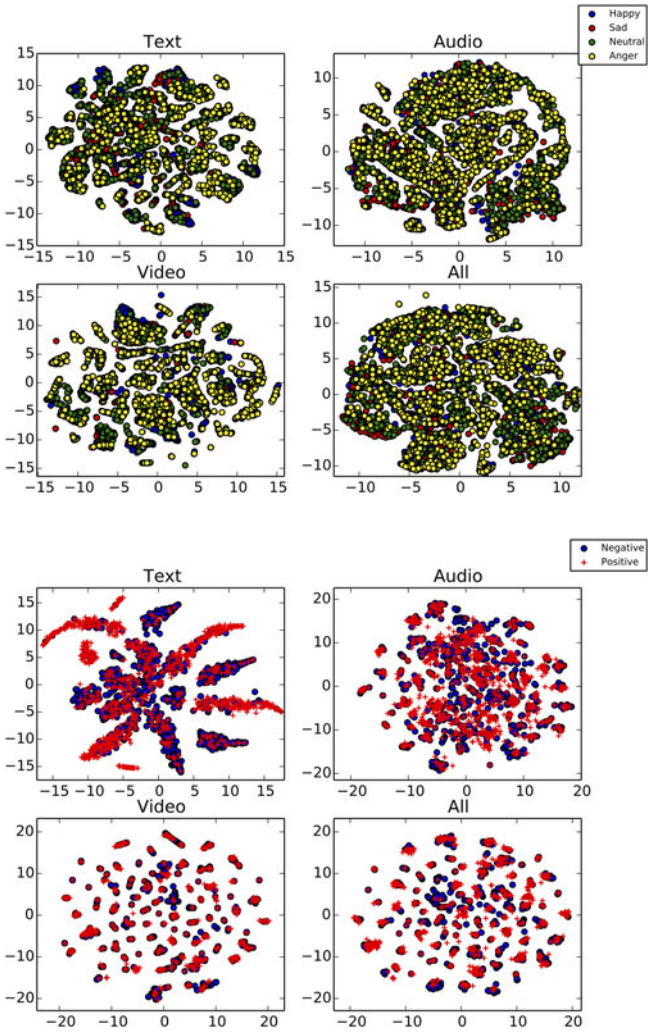

Figure 3. T-SNE 2-D visualization of MOSI and IEMOCAP datasets when unimodal features and multimodal features are used.

multimodal distribution has the least overlap (increase in red and blue visuals, apart from the rest) with sparse distribution aiding the classification process.

CONCLUSION

We have presented useful baselines for multimodal sentiment analysis and multimodal emotion recognition. We also discussed some major aspects of multimodal sentiment analysis problem, such as the performance in the unknown-speaker setting and the 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. C. Busso *et al.*, "IEMOCAP: In teractive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
2. S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. P. Morency, "Multi-level multiple attentions for context-aware multimodal sentiment analysis," in *Proc. Int. Conf. Data Mining*, 2017, pp. 1033–1038.
3. L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multi-modal human emotion/expression recognition," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 366–371.
4. D. Datcu and L. Rothkrantz, "Semantic audio-visual data fusion for automatic emotion recognition," Euromedia, 2008, http://mmi.tudelft.nl/pub/dragos/_datcu_euromedia08.pdf
5. L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proc. IEEE ICICS*, 1997, pp. 397–401.
6. P. Ekman, "Universal facial expressions of emotion," Culture and Personality: Contemporary Readings/Chicago, pp. 151–158, 1974.
7. F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
8. A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using Gaussian mixture models for face and voice," in *Proc. 10th IEEE Int. Symp. ISM*, 2008, pp. 250–257.
9. V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance- level multimodal sentiment analysis," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.
10. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
11. N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl. Based Syst.*, vol. 161, pp. 124–133, 2018.
12. S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user- generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (volume 1: Long papers)*, July 2017, pp. 873–883.
13. S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. Int. Conf. Data Mining*, 2016, pp. 439–448.
14. M. Wollmer *et al.*, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, May/Jun. 2013.
15. A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.
16. Y. Kim, "Convolutional neural networks for sentence classification," arXiv:1408.5882, 2014.

ABOUT THE AUTHORS

Soujanya Poria is a Presidential Postdoctoral Fellow with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Contact him at sporia@ntu.edu.sg.

Navonil Majumder is currently working toward the Ph.D. degree at the Centro de Investigación en Computación (CIC) of the Instituto Politécnico Nacional, Mexico City, Mexico. Contact him at navo@nlp.cic.ipn.mx.

Devamanyu Hazarika is currently working toward the Ph.D. degree at the School of Computing at National University of Singapore, Singapore. Contact him at hazarika@comp.nus.edu.sg.

Erik Cambria is an Assistant Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Contact him at cambria@ntu.edu.sg.

Alexander Gelbukh is a Research Professor with the CIC of the Instituto Politécnico Nacional, Mexico City, Mexico. Contact him at gelbukh@cic.ipn.mx.

Amir Hussain is a Full Professor with the School of Computing, Edinburgh Napier University, Edinburgh, U.K. Contact him at a.hussain@napier.ac.uk.