An Efficient Fusion Mechanism for Multimodal Low-resource Setting

Dushyant Singh Chauhan†, **Asif Ekbal**†, **Pushpak Bhattacharyya**∓
† Department of Computer Science & Engineering
Indian Institute of Technology Patna, India
{1821CS17,asif,pb}@iitp.ac.in

ABSTRACT

The effective fusion of multiple modalities (i.e., text, acoustic, and visual) is a non-trivial task, as these modalities often carry specific and diverse information and do not contribute equally. The fusion of different modalities could even be more challenging under the low-resource setting, where we have fewer samples for training. This paper proposes a multi-representative fusion mechanism that generates diverse fusions with multiple modalities and then chooses the best fusion among them. We also propose an attention mechanism for handling noisy representation that focuses only on contributing representation and ignores the noisy representation. We evaluate our proposed approach on three low-resource multimodal sentiment analysis datasets. Experimental results show the effectiveness of our proposed approach with the accuracies of 59.3%, 83.0%, and 84.1% for the YouTube, MOUD, and ICT-MMMO datasets, respectively.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**; • **Neural networks**;

KEYWORDS

Deep learning, Multi-representative fusion, Low-resource dataset

ACM Reference Format:
Dushyant Singh Chauhan†**, Asif Ekbal**†**, Pushpak Bhattacharyya**∓. 2022. An Efficient Fusion Mechanism for Multimodal Low-resource Setting. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531900

1 INTRODUCTION

Multimodal data provides multiple heterogeneous sources of diverse information. By considering the complex intramodal and intermodal fusions, this diverse information could be used to gain key insights for the various tasks. But, learning these fusions is a fundamentally complex research problem. In recent times, deep neural networks have shown success in achieving good performance for multimodal sentiment, and emotion analysis [Akhtar et al. 2019; Chauhan et al. 2019; Poria et al. 2016, 2017b; Ranganathan et al. 2016; Rosas et al. 2013; Sangwan et al. 2019].

However, multimodal information fusion is not always effective, as different sources often bring their characteristics, and some may contain noise. For example, there might be some disturbances or noise present in a video due to which acoustic features like tone, intensity, energy, pitch, etc., can be affected.

Quite often, different pairs of modalities have semantic inter-dependencies. To understand the semantic inter-dependencies between the different pairs of modalities, the representations that we obtain are fused. If any of the modality representations being fused is noisy, the model will fail to understand the inter-dependencies between modalities. Also, the output of the fusion operation will affect the performance of the model in the subsequent stages.

As we know that not every neuron is helpful for the prediction, as shown in these papers [Chauhan et al. 2020; Mai et al. 2019]. So, motivated by this idea, we propose a multi-representative fusion (MRF) mechanism that first generates diverse representations for each modality and then select the most appropriate representations among them to get the best fusion (i.e., □ in Figure 1). The objective of MRF is to solve the noisy representation problem by leveraging multiple representations of a modality rather than a single representation. Each representation of MRF is unique, which mean no two representation will be exact. Thus, if there is noise in one or more than one modality, then it is quite possible that one of the generated representations is noise invariant or with less noise.

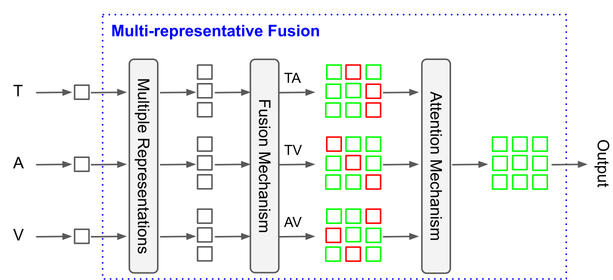

Figure 1: A multimodal architectural view showing the multiple representations and their fusions. In Figure, □ denotes the multimodal fusion corresponding to a noise-free representation while □ denotes the multimodal fusion corresponding to a noisy representation.

We summarize the main contributions as follows: (i). we use convolution filters to generate different and diverse representations of modalities; (ii). we then fuse pairwise modalities with multiple

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531900

representations to get the multiple fusions; (iii). finally, we propose an attention mechanism that only selects the most appropriate fusion, which eventually helps resolve the noise problem and improve the performance. (iv). we present new state-of-the-art systems for three benchmark datasets for sentiment analysis.

2 RELATED WORK

Fusion is the main challenging problem of multimodal data. There are lots of work [Chauhan et al. 2019; Ghosal et al. 2018; Huang et al. 2019; Majumder et al. 2019; Poria et al. 2017a; Zadeh et al. 2017] has already been established on fusion mechanism. Chauhan et al. [Chauhan et al. 2019] exploits the interaction between a pair of modalities through an application of the Intermodal Interaction Module (IIM) that closely follows the concepts of an auto-encoder for the multimodal sentiment and emotion analysis. Poria et al. [Poria et al. 2017a] proposed an LSTM-based framework for sentiment classification that leverages contextual information to capture the inter-dependencies between the utterances. A contextual intermodal attention-based framework for multimodal sentiment classification has been proposed in [Ghosal et al. 2018].

A Deep Multimodal Attentive Fusion, a novel image-text sentiment analysis model, is proposed in [Huang et al. 2019]. In another work [Majumder et al. 2019], authors have proposed a variational autoencoder-based approach for modality fusion and minimized the information loss between unimodal and multimodal representations. Zadeh et al. [Zadeh et al. 2017] proposed a Tensor Fusion Network (TFN) model to learn the intra-modality and inter-modality dynamics of the multimodal.

These fusion mechanisms fail when noisy representation comes into action. In comparison to this existing research, our proposed approach aims to generate diverse representations for each modality and choose the best representations from these. Further, to the best of our knowledge, this is the first attempt to solve multimodal problems through multiple representations

3 MULTI-REPRESENTATIVE FUSION

Our proposed framework aims to leverage the multi-representative fusion mechanism to overcome noise problems from modalities by using multiple representations instead of a single representation. We divide multi-representative fusion into three parts; i) multiple representations, ii) fusion mechanism and iii) attention mechanism.

3.1 Multiple Representations

Given multimodal inputs i.e., text (T), acoustic (A), and visual (V) and generate multiple representations for each modality. We first try to capture the semantic information from each modality by using three separate bi-directional Gated Recurrent Units (Bi-GRU) [Cho et al. 2014] for each modality respectively. Then, we obtain multiple representations for each modality. Thus, to obtain multiple representations, we apply k *convolution filters* on each modality, and these k filters produce k different representations for each modality, which might have quite similar or completely different information, but two of these cannot be the same.

3.2 Fusion Mechanism

After getting k different representations corresponding to each modality, we fuse these multi-representative modalities. We divide this fusion mechanism into two groups, i.e., intramodal fusion (TT, VV, AA) and intermodal fusion (TV, VT, TA, AT, AV, VA). In other terms, we have a total of nine combinations of modalities ($Modality_{Comb}$), i.e., TT, VV, AA, TV, VT, TA, AT, AV, and VA. Fusion between modalities gives us fused scores and fused features, which are as follows;

3.2.1 *Fused Scores:*
The motive of getting fused scores (FS) is to help select the appropriate fusion and help reveal contributing modalities. For example, let's assume there are two modalities x and y $\in \mathbb{R}^{1 \times d}$ where d is the embedding dimension and $x, y \in [T, A, V]$. We multiply both the modalities (i.e., x and y) to extract the FS$\in \mathbb{R}^1$. As, we have k representations *w.r.t.* each modality, we will obtain a matrix of size $k \times k$ by fusing two modalities x and y. Also, there is a total of nine possible $Modality_{Comb}$, so total fused scores (FS)$\in \mathbb{R}^{9 \times k \times k}$.

3.2.2 *Fused Features:*
In contrast, the motive of getting fused features (FF) is to extract meaningful information from the multi-modality, which helps to understand the semantic inter-dependencies between modalities.

We first reshape the modality-wise features to $d \times 1$ and then multiply both the modalities (i.e., x and y) to extract the FF$\in \mathbb{R}^{d \times d}$. Then, we take the sum over each raw to get the normalized features that also reduce the vector size, i.e., FF$\in \mathbb{R}^d$. As, we have k representations *w.r.t.* each modality, we will obtain a matrix of size $k \times k \times d$ by fusing two modalities x and y. Also, there is a total of nine possible $Modality_{Comb}$, so total fused features (FF)$\in \mathbb{R}^{9 \times k \times k \times d}$.

3.3 Attention Mechanism

The fusions that we obtain may be of noisy because it is quite possible that k_x^{th} representation of modality x may not fuse effectively with k_y^{th} representation of modality y, but may instead fuse most effectively with any of $1_y^{st}, 2_y^{nd}, ..., (k-1)_y^{th}$ representation of modality y. So, we apply an attention mechanism to select the best fusions and exclude the noisy fusions among these $9 \times k \times k$ fusions. The detailed steps of the attention mechanism are explained below.

3.3.1 *Softmax:*
We apply Softmax (S) over FS ($\in \mathbb{R}^{9 \times k \times k}$) to compute the probability score[1]. Each value in $9 \times k \times k$ fusions signifies the degree of association between the k_x^{th} and k_y^{th} representation of modality x and y.

3.3.2 *Max and Argmax:*
The max and argmax operations are applied over the softmax matrix (S$\in \mathbb{R}^{9 \times k \times k}$). The max operation is done with the intuition that a particular $Modality_{Comb}$ fusion may have more contribution towards correct prediction rather than other $Modality_{Comb}$. Thus, we get the element-wise max value among all $Modality_{Comb}$ which means we first take k^{th} score from each nine $Modality_{Comb}$ fusion and then take max value among these.

In contrast, the intuition of applying argmax operation is to get the contributing intermodal fusion. We depict the overall pictorial representation of max and argmax operation in Figure 2 where each

[1] Please note that, the sum of each row of softmax matrix is equal to one.

cell o represents a unit value or score. In Figure 2, the first cell of the max matrix shows the maximum values while the argmax matrix gives the contributing $Modality_{Comb}$ (first score is from AV; this is why the argmax value is 8).

Also, max and argmax operations reduce the implementation cost by reducing the matrix size from $9 \times k \times k$ to kk by selecting only contributing $Modality_{Comb}$ instead of taking all.

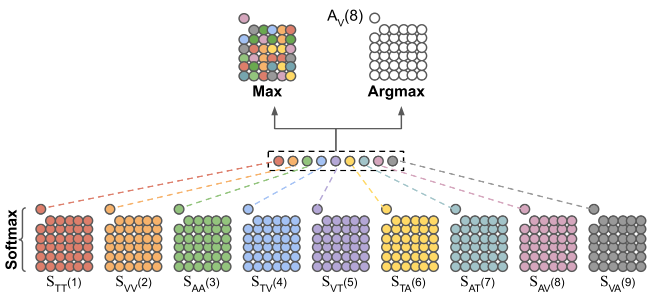

Figure 2: Description of Max and Argmax module. Different colors in the matrix of max describe which combination of modality in contributing the most.

3.3.3 Attentive Features: We have obtained fused features, but we take only those features that contribute to the final prediction and neglect the rest by multiplying with zeros. For attentive features, we use the argmax matrix to extract only the contributing features. Then, a multiplicative gating function [Dhingra et al. 2016] is computed between the attentive features and max matrix (attentive scores) to give priority to attentive features. Finally, we concatenate all the features and pass them through the softmax layer for sentiment classification.

4 DATASETS, EXPERIMENTS, AND ANALYSIS

4.1 Datasets

For the evaluation of our proposed approach, we employ three low-resource multimodal benchmark datasets[2] which are as follows; **YouTube** [Morency et al. 2011]: The dataset contains 269 product review utterances across 47 videos. There are 169, 41, and 59 utterances in training, validation, and test sets. **MOUD** [Pérez-Rosas et al. 2013]: It is a collection of 79 product review Spanish videos where each video consists of multiple utterances labeled with either positive, negative, or neutral sentiment. There are 243 utterances in training, 37 in validation, and 106 in test sets. **ICT-MMMO** [Wöllmer et al. 2013]: It is an extension of the YouTube opinion mining dataset that extends the number of videos from 47 to 340. Each online social review video is annotated with the sentiment class. There are 220, 40, and 80 videos in training, validation, and test sets.

4.2 Experimental Setup

We evaluate our proposed model on three multi-modal datasets: YouTube, MOUD, and ICT-MMMO. For all the three datasets, we perform *grid search* to find the optimal hyper-parameters. Though we push for a generic hyper-parameter configuration for all the datasets, in some cases, a different choice of hyper-parameters has a significant effect on the overall performance. Therefore, we choose different hyper-parameters for different datasets in our experiments. Details of hyper-parameters for different datasets are depicted in Table 1.

Parameters	YouTube	MOUD	ICT-MMMO
Bi-GRU	300N	150N	300N
Dense (FC)	100N	50N	50N
#Filters	8	8	8
#Filter size (H,W)	(1,5)	(1,3)	(1,3)
Stride (H,W)		(1,1)	
Output		Softmax	
Optimizer		Adam (lr=0.001)	
Loss		Cross-entropy	
Batch		8	
Epochs		100	

Table 1: Model configurations

We implement our proposed model in PyTorch, a Python-based deep learning library. As evaluation metric, we employ accuracy and F1-score and *Softmax* as classifier and optimize the *categorical cross entropy* loss. We use *Adam* as an optimizer. Please note that we run experiments using GPUs (GPU: 1080Ti with 11GB, RAM: 256GB).

4.3 Comparative Analysis

We compare the performance of our proposed model against several existing and recent state-of-the-art systems[3]. In particular, we compare with the following systems: Multi-Attention Recurrent Network (MARN) [Zadeh et al. 2018b], Memory Fusion Network (MFN) [Zadeh et al. 2018a], Intermodal Interactive Module (IIM) [Chauhan et al. 2019], Tensor Fusion Network (TFN) [Zadeh et al. 2017], Bi-directional Contextual LSTM (BC-LSTM) [Poria et al. 2017a], Multimodal Factorization Model (MFM) [Tsai et al. 2018].

We show the comparative results in Table 2. The experimental results show the effectiveness of our proposed approach with the accuracies of 59.3%, 83.0%, and 84.1% for the YouTube, MOUD, and ICT-MMMO datasets, respectively. We also perform a statistical significance test (*paired T-test*) between the obtained results and the best score from state-of-the-art systems. We observe that performance improvement in the proposed model over the state-of-the-art is significant with 95% confidence (i.e., p-value< 0.05).

4.4 Ablation Study

To show the efficacy of multi-representative fusion, we perform an ablation study of our proposed model against a basic version without multi-representative fusion, which is called a baseline model. We show the ablation results in Table 3. The ablation study shows the importance of multi-representative fusion with the approximately 5.1, 4.7, and 2.3 performance improvement points over the baseline. We also show a line chart to show the improvement of the proposed MRF over baseline.

[2]These datasets can be accessed through https://github.com/A2Zadeh/CMU-MultimodalSDK.

[3]Please note that we report all the results, which are available for comparison

System	YouTube *F1*	YouTube A^3	MOUD *F1*	MOUD A^2	ICT-MMMO *F1*	ICT-MMMO A^2
MARN†	-	48.3	81.2	81.1	-	-
MFN†	51.6	51.7	80.4	81.1	73.1	73.8
IIM	55.1	55.9	82.0	82.4	81.4	82.7
TFN†	-	-	-	-	72.6	72.5
BC-LSTM†	45.1	-	-	-	-	-
MFM	52.4	53.3	81.7	82.1	79.2	81.3
Proposed	**56.8**	**59.3**	**82.7**	**83.0**	**82.0**	**84.1**
T-test	*0.0005*	*0.0041*	*0.0006*	*0.00003*	*0.031*	*0.040*

Table 2: Comparative results; †Values are taken from [Tsai et al. 2018]. Significance *T*-test (< 0.05) signifies that the obtained results are statistically significant over the existing systems with 95% confidence score. Here, A^3 and A^2 are three-class (negative, neutral, and positive) and two-class (positive and negative) classification accuracies, respectively.

System	YouTube *F1*	YouTube A^3	MOUD *F1*	MOUD A^2	ICT-MMMO *F1*	ICT-MMMO A^2
Baseline	53.3	54.2	77.8	78.3	80.1	81.8
Proposed	**56.8**	**59.3**	**82.7**	**83.0**	**82.0**	**84.1**

Table 3: Ablation results for our proposed model

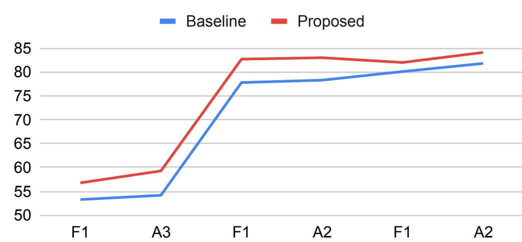

Figure 3: Line chart: baseline vs. proposed MRF

5 ANALYSIS OF MRF

We take a random word from the YouTube dataset and show the argmax matrix (left) corresponding to that word in Table 4.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
f_1	1	2	3	7	8	8	9	5
f_2	9	1	2	5	7	7	8	5
f_3	8	9	1	5	4	5	7	4
f_4	7	8	9	4	3	5	5	3
f_5	5	7	8	3	2	4	5	2
f_6	5	5	7	2	1	3	4	1
f_7	4	5	5	1	9	2	3	9
f_8	3	4	5	5	8	1	2	8

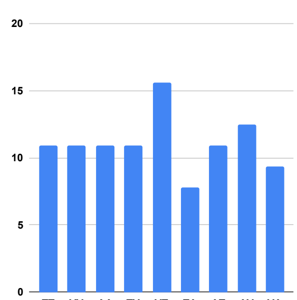

Table 4: Argmax matrix (left) and bar-chart (right) shows the Contribution of Modality$_{\text{Comb}}$

In the argmax matrix, (f_i, f_j) denotes the argmax value, representing the most contributing combination of modalities among all the nine combinations. For example, (f_1, f_5)=8 represents that *AV* is contributing the most when filter-1 of A and filter-5 of V are fused. Similarly, (f_2, f_5)=7 represents that *AT* is contributing the most. Thus, we observe that each filter captures different or diverse information *w.r.t.* modalities. We also show the argmax matrix in the form of a bar-chart (right) to easily show the contribution of each combination of intermodal fusion (in percentage).

We perform a study to justify that MRF is a noise invariant. We make some changes in the dataset in terms of putting zeros[4] instead of actual acoustic embedding for some words in some utterances. We then train the model on this changed dataset, and we get the approximately same accuracy of 59.27% while the proposed accuracy is 59.29%. We observe that model is trying to ignore acoustic modality because of not contributing much in the prediction and focusing on other modalities, i.e., T and V. For the same word as in Table 4, we show argmax and bar-chart in Table 5. The bar-chart (right side in Table 5) clearly shows the reduction in the contribution of acoustic modality because of the noise. This states that if one or more than one noisy modalities are there, then MRF will try to ignore these modalities. This proves the efficacy of our proposed MRF.

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
f_1	1	9	5	1	2	4	5	8
f_2	4	4	8	5	1	2	5	7
f_3	2	4	5	7	9	1	5	4
f_4	5	6	3	1	2	2	4	2
f_5	6	4	5	2	7	4	3	2
f_6	5	1	3	4	6	7	2	1
f_7	2	3	9	4	5	6	1	8
f_8	4	2	1	2	3	4	9	5

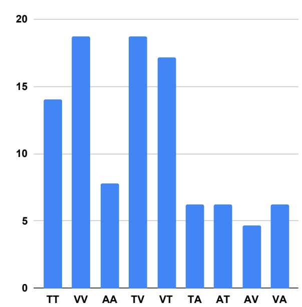

Table 5: Argmax matrix (left) and bar-chart (right) shows the Contribution of Modality$_{\text{Comb}}$

We also perform some other experiments to learn the behavior of MRF. We observe that when one or two modalities are noisy, then MRF works fine. Please note that the term noisy means some utterances are noisy, not all utterances. But when all three modalities are noisy corresponding to a word, which has an important role in the prediction, MRF fails to predict the actual label for the utterances.

6 CONCLUSION

In this paper, we have successfully established the concept of obtaining effective fusions for low-resource multimodal affect analysis. We have proposed a multi-representative fusion mechanism that generates diverse fusions with multiple modalities and then chooses the best fusion among them. We have also proposed an attention mechanism for handling noisy representation that focuses only on contributing representation and ignores the noisy representation. We have evaluated our proposed approach on three multimodal datasets

[4]We replace word-embedding ($\in d$) to zeros and treat zero as a noise.

(i.e., YouTube, MOUD, and ICT-MMMO). Experimental results suggest the effectiveness of our proposed model for sentiment analysis over the existing state-of-the-art systems.

For future work, we would like to apply this multi-representative fusion mechanism to different areas of natural language processing, e.g., machine translation, text summarization, question answering, information retrieval, etc.

ACKNOWLEDGEMENT

Dushyant Singh Chauhan acknowledges the support of the Prime Minister Research Fellowship (PMRF), Govt. of India. Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (Meit/8Y), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

REFERENCES

Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. *arXiv preprint arXiv:1905.05812* (2019). arXiv:1905.05812 [cs.CL]

Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware Interactive Attention for Multi-modal Sentiment and Emotion Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5651–5661.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4351–4360. https://doi.org/10.18653/v1/2020.acl-main.401

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549* (2016).

Deepanway Ghosal, Md Shad Akhtar, Dushyant Singh Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual Inter-modal Attention for Multi-modal Sentiment Analysis. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3454–3466.

Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, and Zhoujun Li. 2019. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems* 167 (2019), 26–37.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 481–492.

Navonil Majumder, Soujanya Poria, Gangeshwar Krishnamurthy, Niyati Chhaya, Rada Mihalcea, and Alexander Gelbukh. 2019. Variational Fusion for Multimodal Sentiment Analysis. *arXiv preprint arXiv:1908.06008* (2019).

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 169–176.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 973–982.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017a. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 873–883.

Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 439–448.

Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. 2017b. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing* 261 (2017), 217–230.

Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems* 28, 3 (2013), 38–45.

Suyash Sangwan, Dushyant Singh Chauhan, Md Akhtar, Asif Ekbal, Pushpak Bhattacharyya, et al. 2019. Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis. In *International Conference on Neural Information Processing*. Springer, 662–669.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems* 28, 3 (2013), 46–53.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, 1103–1114.

A Zadeh, PP Liang, S Poria, P Vij, E Cambria, and LP Morency. 2018b. Multi-attention Recurrent Network for Human Communication Comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-2018)*. New Orleans, USA., 5642 – 5649.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory Fusion Network for Multi-view Sequential Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press.