# An RNN-Based Multimodal Sentiment Analysis: Focusing on Facial Expressions and Multimodal Dynamic Representations

## ABSTRACT

Multimodal sentiment analysis is rapidly gaining popularity due to its potential importance in fields like big data management and human-agent interaction systems. In this paper we present a multimodal opinion classification system fully constructed on deep learning technology. It is RNN-based, thus taking into account the time-dependency of the data. The features used are chosen in a way to allow the system to learn its own internal representation of the data. This was done with the goal to obtain a more accurate system that can generalize better when given a sufficient amount of data, leveraging the deep learning ability to learn features. It was trained and tested on the MOSI dataset which makes it an utterance-level sentiment analysis system. The results presented here are, to the best of our knowledge, the first results obtained on MOSI for such a task (highest accuracy of 84.30%). Additionally, as a validation experiment, we compare and show that our RNN-based system outperforms a CNN-based one (presented in previous work) on the linguistic modality of the MOUD dataset.

## CCS CONCEPTS

•**Computer systems organization** → **Embedded systems;** *Redundancy;* Robotics; •**Networks** → Network reliability;

## KEYWORDS

Multimodal, Sentiment analysis, Deep learning, Recurrent Neural Network, Feature learning

**ACM Reference format:**
. 2017. An RNN-Based Multimodal Sentiment Analysis: Focusing on Facial Expressions and Multimodal Dynamic Representations. In *Proceedings of 19th ACM International Conference on Multimodal Interaction, Glasgow, Scotland, November 2017 (ICMI'17),* 5 pages.
DOI: 10.475/123_4

## 1 INTRODUCTION AND BACKGROUND

Multimodal sentiment analysis has been a growing field this past decade. It can be defined as the automatic analysis or estimation of expressed sentiments (opinions) through different modalities, verbal and non-verbal: written text, spoken words, paralinguistic speech cues, or facial expressions. Its growing popularity is due to the significant contributions it could bring to areas such as big data analysis, user modeling, recommender systems, search engines and Human-Computer Interaction (HCI). There are currently tremendous amounts of unprocessed audiovisual data available on the web and more particularly on social media and review platforms. Automatic processing would enable transforming such raw data into valuable information for businesses. Also, being able to estimate user opinion or sentiment would be beneficial for artificial agents enabling them to derive more accurate understanding of users.

Also, deep learning has proved its efficiency in different areas such as speech processing [19], computer vision [17] and even emotion recognition and sentiment analysis [16]. Deep learning systems also enable feature learning [18], as opposed to feature engineering. In the latter, the data is first processed using signal processing algorithms in a way that suits a machine learning process. In the former, the system learns a representation of the raw data during training. This trend results on one side from deep learning and optimization approaches that are becoming more efficient, and on the other side from the availability of data sets that are large enough for the system to learn on its own an internal representation helpful for solving the target task.

In this paper, we present our work on a multimodal Recurrent Neural Network (RNN)-based system for opinion mining. Facial configurations and motions are represented here either by raw facial landmarks positions extracted from the videos images, along with their temporal derivatives, and also by Action Units (AUs). Action Units are derived from the Facial Action Coding System (FACS) proposed by Ekman and Friesen [5]. Our hypothesis is that given enough training samples, using facial landmarks would perform better thanks to FACS as they constitute rawer data as well as to the feature learning capabilities of our deep learning architecture. The words are used as linguistic features and low level acoustic parameters as audio features.

Background research, as well as our main motivations and contributions are described in more details in section 2. We then present the database used of our work in section 3. We explain the feature extraction per modality in section 4 and detail our model and experiment in section 5. Section 6 describes a comparison experiment carried on as a preliminary study which motivated our choice of RNNs for our system. We finally conclude and give our perspectives for future work in section 7.

## 2 MOTIVATIONS AND CONTRIBUTIONS

Sentiment analysis has been dominated by research on natural language. But multimodal sentiment analysis is a fast growing field that attracts more and more research work [14].

In [13], the authors used audio, visual and linguistic modalities to predict an opinion score at the utterance (sentence) level, with evaluations using the MOUD database. This database is composed of 412 utterances. Using an early fusion approach, they show a

*ICMI'17, Glasgow, Scotland*

DOI: 10.475/123_4

10.5% error rate reduction using all modalities compared to the best performing system trained on a single modality (linguistic). In [15], the authors compared early and late fusion techniques for sentiment classification, also considering linguistic, audio and visual modalities, and evaluating their approach on the Youtube dataset [12]. This dataset consists of 47 videos from the social media web site YouTube, in which people are talking about different topics. The authors show an accuracy of 80%, around 20% more than the previous state of the art results on that data set. The linguistic features were extracted using the sentic computing paradigm, 6373 acoustic features for the audio modality using the OpenEAR toolkit [6] and distances calculated from landmarks extracted from the video images were processed for the visual features.

In [16], the same authors propose an MKL-based multimodal sentiment analysis system. The full system is trained on the MOUD dataset and tested on the MOUD, Youtube and ICT-MMMO datasets [21]. The latter is composed of 370 Youtube review videos and the opinion scores were not annotated on the utterance level. The accuracies obtained were all high – in particular 85.3% on the ICT-MMMO – using the visual sentiment model trained on the MOUD dataset. Visual features were extracted using a combination of Convolutional Neural Network (CNN) and an RNN. The raw images were fed to the system. The linguistic features were extracted using a CNN on the text. The acoustic features using OpenEAR in the same way as in the previous paper. We use the MOSI dataset [22] for which opinion scores are annotated at the utterance level. These are 2199 subjective sentences extracted from the ICT-MMMO database. To the best of our knowledge, few work were presented to this database due to it being relatively new. In [2], the others propose opinion scoring results using a CNN architecture to extract linguistic and visual features and an Support Vector Machine (SVM) for fusion on MOSI. The highest F-scores on speaker dependent and independent experiments are reported to be 73.55% and 76.66% respectively after fusing all modalities. In [20] on the speaker independent experiment where the authors prose a Select-Additive Learning system that aims at improving the generalization. They show a highest accuracy score of 0.732 using the text data only and 0.73 using all modalities.

Our contribution in this paper is to propose a multimodal sentiment analysis system which:

- is fully based on deep learning technology;
- integrates temporal information for each modality along a complete utterance in a end-to-end fashion;
- investigating the use of rawer features, enabling feature learning by the deep learning architecture;

The system is built using three RNNs (see Section 5), one to create a representation of each modality, as well as a fusion layer for reaching a multimodal decision. RNN proved to be efficient for time sequence modeling, feature extraction, and classification.

The main modalities concerned with feature learning for this work are the visual and linguistic modalities. For the visual modality, we will use and compare facial landmarks (along with their first temporal derivatives) and AUs. Using AUs, like in [13] or distances calculated from landmarks, as seen in [15] possibly limits the feature learning potential of such systems. On the other hand, using completely raw images as input, as was done in [16], remains a challenge given the moderate size of training datasets available for multimodal opinion mining research. In fact, along with the facial expressions, raw images contain other information (such as the background, the color contrasts, etc...) which makes it more complex for the system to model accurately the data with respect to the target task. Using raw facial landmarks seemed a good compromise to us. We will show that it indeed yields to performances improvements. For the linguistic features, we propose to use embedded vectors randomly initialized and let the system learn representative features.

In this work, the main goal is to show the contribution the features we propose to this task. For all the previously mentioned work, the opinion scoring estimation problem was reduced to a binary classification problem of positive and negative sentiments. The same will be done in this work.

We also present in this paper our preliminary analysis showing results of RNN outperforming CNN on the linguistic modality of the MOUD database. This experiment motivated the choice of RNNs for our system.

# 3 DATASET USED

The dataset used in this work is the MOSI dataset. As mentioned previously, this dataset is constituted of 2199 subjective sentences/videos extracted from the ICT-MMMO database as described in [22]. The sentiment intensity of each sentence was annotated by 5 annotators on a linear scale of integers from -3 to +3. The values correspond to the following labels: strongly positive (+3), positive(+2), weakly positive (+1), neutral (0), weakly negative (-1), negative(-2), strongly negative (-3). The annotators were also given the choice of "uncertain". A mean value for each sentence was then computed. For the purpose of this study and similarly to the previous related work, we consider the problem to be classification problem of positive versus negative sentiment. So, all the positive values were considered as belonging to the positive class and all the negative values to the negative class. The sentences with neutral opinions (96 sentences) were all discarded. One of the videos was also removed due to a bad segmentation which leaved us with 2102 videos.

# 4 FEATURE EXTRACTION

## 4.1 Linguistic Modality Features

Every sentence in the dataset is already tokenized and punctuationless hence no preprocessing is required. We discard every sentence with more than 65 words. Our filtered dataset is now of size 2096. We don't use any pre-trained embeddings. Each word of the sentence $x_i$ is is a row index in a lookup or word embedding matrix $M_x \in \mathbb{R}^{|V_x| \times E_x}$ where $|V_x|$ is the vocabulary size and $E_x$, as previously mentioned, the word embedding size. The embedding matrix $M_x$ is trained along with the model.

## 4.2 Audio Modality Features

The audio features were extracted using the OpenSMILE toolkit [7]. The list of the features are the pitch, voicing probability, voice, loudness/intensity, energy and 12 order MFCCs features with their derivatives and double derivatives. The features were extracted using a 25 ms window width with a 10 ms shit. This leaves us with 43 features. We sample the features to a maximum of 65 samples per sentences. The sampling is done uniformly and we make sure that at least one sample is picked for each word.

### 4.3 Visual Modality Features

The visual features were extracted using the OpenFace toolkit [1]. This toolkit is able to recognize a subset of 17 AUs. It gives, for each AU and each frame of the video, whether or not the AU was detected on the face and the intensity of the AU. These intensity estimation and the AU detections are given separately. For our system, we use only the estimated AU intensities where the corresponding AU was detected. The other intensities are put to 0. OpenFace is also able to extract 68 facial landmarks. Two types or coordinates are given, the 2D coordinates in pixel coordinates and 3D coordinates in millimeters. For this work, since we intended to feed the data directly to the system and since the Z coordinate in the 3D landmarks had 100 times the values of the other two coordinates, we chose the 2D landmarks. The derivatives from the 2D landmarks coordinates were then computed. We final obtain 289 visual features (68 2D landmarks, 68 Δ2D landmarks and 17 AUs). In order to have the same range of values between all the visual features, the 2D landmarks' values were divided by 100. A sampling is done in the same fashion than for audio.

## 5 RNN MULTIMODAL SYSTEM

Given a sequence $X = (x_1, x_2, \ldots, x_M)$ where each $x_i \in \mathbb{R}^{E_X}$ is a component of this sequence, the neural network directly models the probability of the sentiment polarity according to the sequence X. The network consists of one bidirectional RNN encoder and one Multi-Layer Perceptron (MLP). The encoder computes a representation $C = (c_1, c_2, \ldots, c_N), c_i \in \mathbb{R}^{E_C}$ for each sequence and the MLP outputs the score for two classes (positive or negative sentiment).

More formally, at every time-step of the sequence, the forward RNN $\overrightarrow{\Psi}_{\text{enc}}$ computes a sequence of annotations $(\overrightarrow{h}_1, \overrightarrow{h}_2, \ldots, \overrightarrow{h}_M)$ called hidden states by iterating the following equation:

$$h_t = \varphi(W_x x_t + W_h h_{t-1})$$

where $\varphi$ is a non-linear function. The backward RNN $\overleftarrow{\Psi}_{\text{enc}}$ reads the input sequence in reverse order and produces a set of annotations $(\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_M)$. The bidirectional RNN yields to the MLP the final sequence representation $C = [\overrightarrow{h}_M; \overleftarrow{h}_M]$. We chose the Gated Recurrent Unit (GRU) [3, 4] as non-linearity function $\varphi$. This type of hidden unit has been motivated by the Long Short-Term Memory (LSTM) unit (GRU contains a forget and reset gate too) but is much simpler for computation and implementation. Even though both cells lead to similar results, we found the GRU to be more efficient and better suited for our experiments.

The input layer of the MLP takes as input the representation $C$. The continuous state $h_j$ of an input neuron $j$ is computed as a weighted sum over every $c_i$ and with the bias $b_j$. The output of the neuron is the result of a non-linearity over the hidden state $h_j$:

$$h_j = \sum_{i=1}^{n} w_{ij} c_i + b_j = w_j^T + b_j$$

$$y_j = \varphi(h_j)$$

For the next iteration, the following hidden layers takes as input the output of the preceding one. The last layer has two outputs in order to compute the probabilities over the positive and negative opinion. For our experiments, we choose ReLU for our MLP nonlinear function $\varphi$. The whole model is trained end-to-end as shown in figure 1.

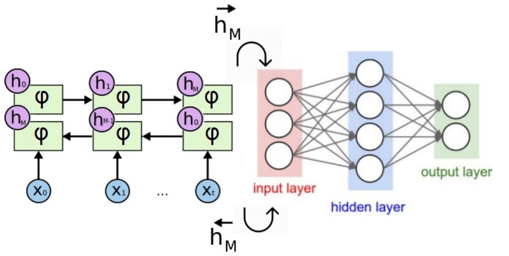

**Figure 1: End to end modality training**

### 5.1 Mono-modality

For each modality, we use a GRU layer size $E_C$ of 512. Every final representation $C$ is then a vector of 1024. This vector is fed through the Fully Connected Neural Network (FC) which consists of four layers of size $[1024 \times 512, 512 \times 256, 256 \times 128, 128 \times 2]$. All recurrent matrices are random orthogonal and bias vectors are all initialized to zero. We use the Adam optimizer [11] with an L2 regularization of $\delta = 10^{-5}$. The learning rate vary from modality to modality. We use a mini-batch size of 100. We also apply dropout of 0.2 on the cell state and the cell output at every time-step.

**Linguistic Modality** We pick an embedding size $E_X$ of 620 initialized by sampling from a Gaussian $\mathcal{N}(0, 0.01^2)$ and a learning rate of 0.0001.

**Audio Modality** The audio frames have an embedding size $E_X$ of 172 and a learning rate of 0.0003.

**Video Modality** The video frames have an embedding size $E_X$ of 289 – 17 Action Units (AU) , 2D Landmarks (L) and 2D Landmarks derivatives (ΔL) – and a learning rate of 0.0003.

All scores presented in the following table are the result of a 10-fold cross validation.

From the table it is obvious that the linguistic cue discriminates the positive and negative sentences the best. Concerning the visual cue, we can see that adding the RL to the AUs increase the accuracy by 2.37%. The choosing features that allow the deep learning model to learn its own representations of the data with respect to the target task should thus be considered in future work. The best score for the visual cue was obtained when combining all three features. This means that the landmarks coordinates derivatives help improve the visual results. A reason may be the sampling on the frames. Indeed, the RNN and FC layers are not able to compute the derivatives of the coordinates themselves. Having the information as a feature may help the model to learn better. In the future, using a single system that learns a common representation

| Modality | Test Accuracy |
|---|---|
| Linguistic | **70.81 %** |
| Audio | **60.76 %** |
| Video | |
| [AU] | 48.33 % |
| [AU, L] | 50.70 % |
| [L, ΔL] | 52.53 % |
| [AU, L, ΔL] | **54.54 %** |

**Table 1: Results on the MOSI dataset modality-wise.**

of the data and thus discarding the frame sampling step, might improve the results without the need of the ΔL features

## 5.2 Multi-modality

For every sentence in the dataset, we extract its representation for each modality and concatenate them in a single vector. A sentence is now embedded in a vector of size 3096. For the multimodality classification, we only use our fully-connected neural network but with input size 3096. We report scores in the following table:

| Modality | Test set Accuracy |
|---|---|
| Linguistic+Audio | **82.29 %** |
| Linguistic+Video | **81.80 %** |
| Linguistic+Audio+Video | **84.30 %** |

**Table 2: Results on the MOSI dataset with multimodality**

The main conclusion that can be drawn from these results is that combining the modalities, increase the the accuracies from 10 to almost 14%.

# 6 RNN VS CNN ON LINGUISTIC MODALITY

In [16], the authors presented interesting results using CNNs on the linguistic cue of the MOUD dataset before and after translation of the sentences from Spanish to English. As a study preliminary to this work, a comparison was made between an RNN system and the same CNN as the one described in [16] before the translation in English (case of non-translated sentences). CNNs have recently achieved strong performance on the practically important task of sentence classification in [8–10]. Our CNN was built similarly as in these papers. Every word of the sentences are embedded in a vector $w_i \in \mathbb{R}^E$. A sentence $s$ is now represented as the concatenation of its words $s = w_1 \oplus w_2 \oplus \ldots \oplus w_n$. If the length of a given sentence is $m$, $d$ the number of words in a sentence, then the sentence matrix dimensionality is $m \times d$. So the matrix can have the same dimension $d$ for every sentence, we use zero-padding [10], on shorter sentences. We can now treat the sentence matrix as an 'image', and perform convolution on it via linear filters. Because rows represent discrete symbols (words), we use filters with widths equal to the dimensionality $E$ of the word vector. The only variable is the height of the filter i.e., the number of adjacent words.

Our CNN consist of 4 layers :

- Input layer taking the sentence matrix
- Convolution layer with two filters of height size 3 and 4 computing two feature maps of dimensionality 78, and one filter of height 2 computing two feature maps of dimensionality 100
- Max-pool layer performing an 1-max-pool.
- Output softmax layer of 2 neurons for classification.

RNN was built similarly as above.

| Method | Test set | |
|---|---|---|
| | Accuracy | F-score |
| Our CNN + MLP | 65.91 % | |
| CNN + SVM [16] | 68.56 % | |
| BoW + SVM [13] | 70.94 % | |
| CNN + SVM [2] | | 48.40 % |
| Our RNN + MLP | **72.72 %** | **70.45 %** |

**Table 3: Results on the MOUD dataset**

In addition, RNNs can handle arbitrary input lengths making the process faster, whereas CNNs take a fixed size input (all sequences need to be padded with a specific token in order to have the same length as previously mentioned). Figure 2 illustrates the difference of performance. For both figures, CNN and RNN have the same batch size and their learning rate has been optimally and empirically chosen for each. We also would like to point out flexibility of the RNN models : we included the neutral opinion in the dataset (making it a 3 classes classification), increased the layer size $E_C$ from 512 to 1024 and obtained an accuracy 71.42 %.

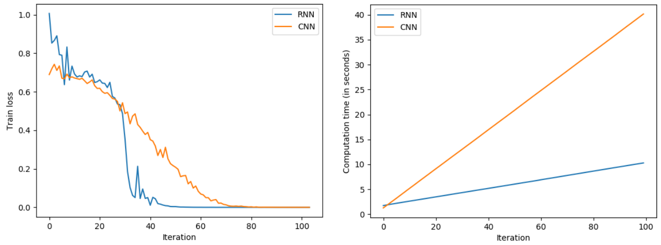

**Figure 2: Left: The CNN training converges slower to the optimum - Right: The CNN is consequently longer to train.**

# 7 CONCLUSIONS AND FUTURE WORK

In this work we presented an RNN-based system for opinion classification. We exposed first results on the MOSI database with the highest accuracy obtained when combining all modalities. We also investigate the use of facial expression descriptors that are landmarks along with their derivatives and AUs. In this work, different models have learned the representation of each modality separately. They were then combined and connected to an FC network for multimodal classification. A highest accuracy of 84.30% was obtained when combining all modalities In future and ideally, an end-to-end system should be built with the goal to learn a representation of the data from all modality at once.

## REFERENCES

[1] T. Baltrusaitis, P. Robinson, and L. P. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10. https://doi.org/10.1109/WACV.2016.7477553

[2] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and R.B.V. Subramanyam. 2017. Benchmarking Multimodal Sentiment Analysis. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*.

[3] Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. http://www.aclweb.org/anthology/D14-1179

[4] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. *Empirical evaluation of gated recurrent neural networks on sequence modeling*.

[5] P. Ekman and W. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.

[6] F. Eyben, M. Wöllmer, and B. Schuller. 2009. OpenEAR: Introducing the munich open-source emotion and affect recognition toolkit. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. 1–6. https://doi.org/10.1109/ACII.2009.5349350

[7] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[8] Rie Johnson and Tong Zhang. 2015. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. 103–112.

[9] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A Convolutional Neural Network for Modelling Sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. 655–665.

[10] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 1746–1751.

[11] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] 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 (ICMI '11)*. ACM, New York, NY, USA, 169–176. https://doi.org/10.1145/2070481.2070509

[13] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-Level Multimodal Sentiment Analysis.. In *ACL (1)*. 973–982.

[14] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98 – 125. https://doi.org/10.1016/j.inffus.2017.02.003

[15] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174, Part A (2016), 50 – 59. https://doi.org/10.1016/j.neucom.2015.01.095

[16] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain. 2016. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 439–448. https://doi.org/10.1109/ICDM.2016.0055

[17] Joseph Redmon and Ali Farhadi. 2016. YOLO9000: Better, Faster, Stronger. *CoRR* abs/1612.08242 (2016). http://arxiv.org/abs/1612.08242

[18] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, and year=2016 pages=5200-5204 Björn W. Schuller and Stefanos Zafeiriou, journal=2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. (????).

[19] Aron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. In *Arxiv*. https://arxiv.org/abs/1609.03499

[20] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2016. Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis. *CoRR* abs/1609.05244 (2016). http://arxiv.org/abs/1609.05244

[21] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. P. Morency. 2013. YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems* 28, 3 (May 2013), 46–53. https://doi.org/10.1109/MIS.2013.34

[22] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems* 31, 6 (Nov 2016), 82–88. https://doi.org/10.1109/MIS.2016.94