Multimodal Relational Tensor Network for Sentiment and Emotion Classification

Saurav Sahay **Shachi H Kumar** **Rui Xia** **Jonathan Huang** **Lama Nachman**
Anticipatory Computing Lab
Intel Labs
{saurav.sahay, shachi.h.kumar, rui.xia, Jonathan.Huang, lama.nachman}
@intel.com

Abstract

Understanding Affect from video segments has brought researchers from the language, audio and video domains together. Most of the current multimodal research in this area deals with various techniques to fuse the modalities, and mostly treat the segments of a video independently. Motivated by the work of (Zadeh et al., 2017) and (Poria et al., 2017), we present Relational Tensor Network architecture where we use the inter-modal interactions within a segment and also consider the sequence of segments in a video to model the inter-segment inter-modal interactions. We also generate rich representations of text and audio modalities by leveraging richer audio and linguistic context alongwith fusing fine-grained knowledge based polarity scores from text. We present the results of our model on CMU-MOSEI dataset and show that our model outperforms many baselines and state of the art methods for sentiment classification and emotion recognition.

1 Introduction

Sentiment Analysis is broadly defined as the computational study of subjective elements such as opinions, attitudes, and emotions towards other objects or persons. Sentiments attach to modalities such as text, audio and video at different levels of granularity and are useful in deriving social insights about various entities such as movies, products, persons or organizations. Emotion Understanding is another closely related field that commonly deals with analysis of audio, video, and other sensory signals for getting psychological and behavioral insights about an individual's mental state. Emotions are defined as brief organically synchronized evaluations of major events whereas sentiments on the other hand are considered as more enduring beliefs and dispositions towards objects or persons (Scherer, 1984). The field of Emotion Understanding has rich literature with many interesting models of understanding (Plutchik, 2001) (Ekman, 2009) (Posner et al., 2005).

In this work, we explore methods that combine various unimodal techniques for classification alongwith multimodal techniques for fusion of cross modal interactions to perform sentiment analysis and emotion understanding. We develop and test our approaches on the CMU-MOSEI dataset (Zadeh et al., 2018d) as part of the ACL Multimodal Emotion Recognition grand challenge. CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset is a newly released large dataset of multimodal sentiment analysis and emotion recognition on YouTube video segments. The dataset contains more than 23,500 sentence utterance videos from more than 1000 online YouTube speakers. The dataset has several interesting properties such as being gender balanced, containing various topics and monologue videos from people with different personality traits. The videos are manually transcribed and properly punctuated. Since the dataset comprises of natural audio-visual opinionated expressions of the speakers, it provides an excellent testbed for research in emotion and sentiment understanding. The videos are cut into continuous segments and the segments are annotated with 7 point scale sentiment labels and 4 point scale emotion categories corresponding to the Eckman'a 6 basic emotion classes (EKMAN, 2002). The opinionated expressions in the segments contain visual cues, audio variations in signal as well textual expressions showing various subtle and non-obvious interactions across the modalities for both sentiment and emotion classification.

What differentiates our work from existing lit-

Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pages 20–27,
Melbourne, Australia July 20, 2018. ©2018 Association for Computational Linguistics

erature is (i) application of a novel cross modal fusion technique across the temporal segments of the multimodal channel (ii) use of rich shallow semantic domain knowledge that include a large number of psycholinguistic features and resources for sentiment and emotion classification and (iii) extraction of emotion aware acoustic phoneme level features using a novel method and architecture.

Our unimodal research focus in this paper is an exploration of speech sentiment and emotion recognition using various text dependent and text independent techniques. On the text modality experiments, we've explored (i) fusion of Lexicons as additional input features (ii) fusion of polarity discriminating lexico-syntactic fine-grained scores as additional input features (iii) fusion of rich contextualized embeddings as additional input features to the classification pipeline. On audio modality, we've used a novel pipeline to generate the iVectors and Phoneme level utterance features. For fusion of multimodal information, we have explored techniques that leverage intramodal and inter-modal dynamics and fused them together in a novel Relational Tensor Network architecture.

2 Related Work

Sentiment Analysis has received a lot of prior attention in Movie reviews and Product reviews domain and is an established field of research in NLP (Liu, 2010) (Pang and Lee, 2008). However, this hasn't been widely researched in conversational multimodal audio-visual and textual context for continuous recognition of sentiments and emotions. (Kaushik et al., 2013) perform sentiment extraction on natural audio streams using ASR on Youtube videos. They use a maximum entropy classifier and do not use any lexicon features or do any domain adaptation. Multimodal Affect recognition has lately gained a lot of popularity with release of multiple datasets and approaches (Zadeh et al., 2018c) (Zadeh et al., 2018b). (Zadeh et al., 2017) present a tensor fusion technique to generate a fused representation of the individual modalities. Most of these techniques treat the segments of a video independently and ignore the temporal relations and interactions between the segments of a video. (Poria et al., 2017) present an LSTM based network architecture that leverages the context or the temporal interactions between neighboring segments of a video by concatenation of cross modal features across the segments. For acoustic emotion recognition, one of the most successful system is based on the super-segmental acoustic features which is extracted by applying multiple functions on frame-level features. These features have been adopted as the baseline system in many acoustic emotion challenges (Schuller et al., 2016) (Valstar et al., 2016) (Dhall et al., 2013). Deep learning techniques have also been used in acoustic emotion recognition system in recent years. In (Neumann and Vu, 2017), convolutional neural network (CNN) is applied on the frame-level feature. In (Tao and Liu, 2017), recurrent neural network (RNN) is used to model the temporal information for emotion recognition system.

3 Model Description

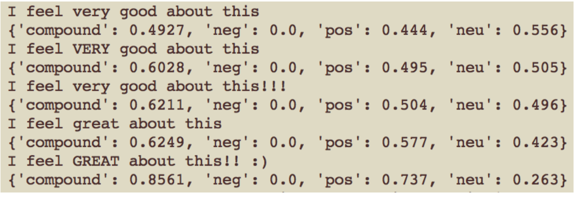
This work brings together techniques for various modality specific feature extraction methods and fusion of information from different modalities for Sentiment and Emotion Classification. The grand challenge dataset comes with modality specific features for text, audio and images as a part of the CMU Multimodal Data SDK (Zadeh et al., 2018a). The text features are based on Glove embeddings (Pennington et al., 2014), audio features are based on COVAREP (Degottex et al., 2014)and the visual features based on FACET (Baltruaitis et al., 2016) visual feature extraction libraries. We extracted various additional features for text and audio modalities as described in the following sections.

3.1 Text

Several traditional methods have been developed in Sentiment Analysis technology for decades before the recent advances in deep learning that primarily rely on methods for word vector representation and automated feature discovery from snippets. We look at modeling some of the traditional methods and features in the deep pipeline and study the impact of these on the classifiers. Below, we describe a couple of traditional knowledge based resources alongwith some recent deep representations that we have fused together in our pipeline.

3.1.1 Lexico-syntactic Rule based features

Text is processed to intrinsically understand the deeper lexico-syntactic patterns to relate them



```
I feel very good about this
{'compound': 0.4927, 'neg': 0.0, 'pos': 0.444, 'neu': 0.556}
I feel VERY good about this
{'compound': 0.6028, 'neg': 0.0, 'pos': 0.495, 'neu': 0.505}
I feel very good about this!!!
{'compound': 0.6211, 'neg': 0.0, 'pos': 0.504, 'neu': 0.496}
I feel great about this
{'compound': 0.6249, 'neg': 0.0, 'pos': 0.577, 'neu': 0.423}
I feel GREAT about this!! :)
{'compound': 0.8561, 'neg': 0.0, 'pos': 0.737, 'neu': 0.263}
```

Figure 1: Sentiment Analyzer

with world knowledge to extract meaningful inferences such as sentiments and emotions. We have explored the use of VADER rules (Hutto and Gilbert, 2014) for sentiment and emotion induction. VADER is a simple and fast rule-based model for general sentiment analysis. It utilizes a human-validated general sentiment lexicon and general rules related to grammar and syntax. The goal of this work is to capture generalizable rules and heuristics associated with grammatical and syntactical cues people use to assess sentiment intensity in text. We can clearly see from Figure 1 how this system can differentiate emphasis, intensity and non-linguistic cues from utterances. Deep learning based systems today fail to capture such systematic nuances deterministically.

3.1.2 Sentiment Lexicons

Lexicons consists of maps of key-value pairs, where the key is a word and the value is a list of sentiment scores for that word (e.g., probabilities of the word in positive, neutral, and negative contexts). The scores have different ranges for showing very negative to very positive sentiments. Lexicon embeddings are sparse signals derived by taking the normalized scores from multiple sources of lexicon datasets. The simplest method of blending a lexicon embedding into its corresponding word embedding is to append it to the end of the word embedding. The General Inquirer(GI) (Stone et al., 1966) is a text analysis application with one of the oldest manually constructed lexicons still in widespread use. It contains 11000 words in 183 different psycholinguistic categories. We have used the lexicon based General Inquirer classes that are divided into groups such as valence, semantic dimensions, cognitive orientation, institutional context, motivation related words, classes of Power, Respect, Affection, Wealth, Well-being, Enlightenment, Skill, etc.. (Shin et al., 2017) who originally explored this work in depth show that lexicon embeddings allow building high-performing models with much smaller word embeddings.

3.1.3 Contextualized Language Embeddings

In contrast to the above two features, we have also looked at recent developments in contextualized deep word vector representations and how they can help with sentiment and emotion classification. These word vectors are learned functions of the internal states of a deep bidirectional language model, which is pretrained on a large text corpus. The additional language modeling views to vector generation process results in high quality representations (Peters et al., 2018). These word vector representations try to model the complex characteristics of word use along with how these uses vary across linguistic contexts (i.e., to model polysemy). We have used ELMo that learns a linear combination of the vectors stacked above each input word for each end task, which improves performance over just using the top LSTM layer (McCann et al., 2017) . Unlike most widely used word embeddings (Pennington et al., 2014), ELMo word representations are functions of the entire input sentence

3.2 Audio Features

For this task, three different kinds of features were applied. The first one is the feature set extracted by using COVAREP which is provided by the challenge. It includes multiple kinds of frame level acoustic features, such as Mel-frequency Cepstral Coefficients (MFCCs), energy and etc. More details are described in (Gusfield, 1997). Along with COVAREP features, we proposed two additional feature-sets, i-vector features and phoneme level features. Following two sections will discuss details about proposed feature-sets.

3.2.1 I-vector Features

The previous studies (Xia and Liu, 2016) (Tao et al., 2018) have shown that i-vector feature can benefit acoustic emotion recognition system. I-vector modeling is a technique to map the high dimensional Gaussian Mixture Model (GMM) supervector space (generated by concatenating the mean of the mixtures from GMM) to low dimensional space called total variability space T.

Give an utterance u, x_t^u which represents t-th frame of utterance u. Audio frame x_t^u is generated by the following distribution:

$$x_t^u \sim \sum_c p(c|x_t^u)\mathcal{N}(m_c + Tw^u, \Sigma_c) \qquad (1)$$

where $p(c|x_t^u)$ is the posterior probability of c-th Gaussian in Universal Background Model (UBM), m_c and Σ_c represent the means and covariance of c-th Gaussian and w^u is the latent i-vector for utterance u. EM algorithm introduced in is applied to iteratively train T. Note that UBM is a GMM which trained with a large corpus.

In (Lei et al., 2014), the phonetically-aware DNN is used to replace the traditional UBM in the framework of i-vector training which showed significant improvements on the speaker identification task. The phonetically-aware DNN is the network for the acoustic model of the Automatic Speech Recognition (ASR) system. It is trained for recognizing the tri-phone state. Compared to the traditional trained UBM, the DNN from ASR represents the feature space constrained on predefined tri-phone states. The posterior probability as the output of this DNN is directly used as the $p(c|x_t^u)$ in Equation 1. In this work, ASR and i-vector extractor are pre-trained on Librispeech dataset (Panayotov et al., 2015) with Kaldi (Povey et al., 2011). We used 960 hours speech data from Librispeech to train DNN-HMM ASR and 460 clean data for i-vector extractor. To avoid overfitting, the dimensionality of i-vector is set as 100. We also tried larger i-vector dimensions but the i-vector with larger dimensions show similar performance compared to the i-vector system with size 100 dimensionality.

3.2.2 Phoneme Level Features

The phoneme related information have also been applied in emotion recognition system. Phoneme-dependent hidden Markov model (HMM) was proposed for emotion recognition system in (Lee et al., 2004). (Bitouk et al., 2010) proposed to extract class-level spectral features on three types of phoneme. Unlike most other work that need accurate alignment, we propose to use the statistics of posterior probability of phoneme on utterance level. The following steps are used to extract phoneme level features:

- Step One: Each frame x_t^u in utterance u is been feed into DNN pre-trained for ASR. The output is a numeric vector consisting of $p(s_i|x_t^u, DNN)$, which corresponding to posterior probability of triphone state s_i. The number of triphone state is dependent on the decision tree algorithm in the ASR system.

- Step Two: Mapping the tri-phone state s_i into monophone. The number of the triphone state is huge. For emotion recognition system, it is not necessary to know information in such fine-grained unit. Instead, we map the triphone state into monophone level by disregarding left and right phone in the triphone structure. The mapping function is: $F_{map}(s_i) = m_j$ where s_i is the tri-phone state and m_j represents corresponding monophone. For example, $F_{map}(r-ae-n) = ae$.

- Step Three: Calculating the statistics of posterior probability of phoneme on utterance level. Given x_t^u, for each cluster m_j, we sum up the posterior probability $p(s_i|x_t^u, DNN)$ once s_i belongs to cluster m_j.

$$P_{m_j}(x_t^u) = \sum_{F_{map}(s_i)=m_j} p(s_i|x_t^u, DNN) \tag{2}$$

  It generates a vector in the length of number of monophone for each frame in utterance u. In order to obtain a fixed dimensional feature for each utterance with variable length, statistics functionals, mean and standard deviation are applied on $P_M(X)$.

For each utterance, the generated feature set is a fixed dimensional vector. Based on the trained DNN-HMM ASR system, the number of tri-phone states and monophone ends up in 5672 and 58 respectively. After mapping and feature extraction, the dimensionality of the phoneme level features is 106.

4 Network Architectures

The models described here are based on a recurrent architecture and use different fusion strategies such as concatenation or tensor fusion across all modalities as well as across all segments of the video. We have integrated ideas from the two distinct approaches to jointly leverage multimodal fusion across modalities and across temporal segments and developed our Multimodal Relational Tensor Network.

4.1 Tensor Fusion Network

TFN consists of a Tensor Fusion Layer that explicitly models the unimodal, bimodal and trimodal inter-modal interactions using a 3-fold Cartesian product from modality embeddings. Most common deep learning approach for fusion of signals is a algebraic Merge operation where the operators are generally a linear concatenation of features or a sum. TFN, on the other hand, tries to disentangle unimodal, bimodal and trimodal dynamics by modeling each of them explicitly. Tensor Fusion is defined as the three-fold Cartesian product amongst the modalities with an extra constant '1' added to the dimension. The extra constant dimension with value '1' analytically generates all the multimodal dynamics followed with vector dot operations. This definition is mathematically equivalent to a differentiable outer product between the modalities. This operation results is a very large number of dimensions in the merged layer and therefore can realistically be applied to problems where the interaction space is not too large.

4.2 Contextual LSTM

Utterances in a video maintain a continuous sequence and work like state machines following a certain path before changing courses. Statistical Sequence classification techniques are applied in the classification of each member of the sequence by modeling the dependence on the other members of the sequence. Human reactions are also generally continuous and maintain a certain state in the sequence before jumping to another state. In particular, it has been seen that, when classifying one utterance, other utterances can provide important contextual information. This natural phenomenon directly maps to methods such as recurrent network approaches and sequence models to capture the dependencies between the segments. We reuse this idea to capture this flow of informational triggers across utterances using an LSTM-based recurrent neural network (RNN).

4.3 Relational Tensor Network

While TFN has been used to model the modality interactions within a video segment, we extend that approach to apply it to the contextual stream of segments. There are two ways we can apply a tensor fusion (by tensor fusion, we specifically refer to the cartesian product operation between the modalities with an extra '1' input to model the inter-modal interaction) across modalities across the streams. The first approach is to apply a tensor fusion across all modality features of all segments for all the modalities. This approach ideally captures all possible cross-dynamics (unimodal, bimodal, trimodal) amongst all possible features of all the video segments. The main issue with this approach is that we run into an exponential growth in the feature space with every modality added in the interaction. The cartesian product further creates multiple outer products for bimodal and trimodal interactions. Even with a small number of features for this approach, our network had about 10s of billions of parameters and this would not be a feasible approach unless deployed on a massive infrastructure. This approach does not require the use of LSTMs as used in the Contextual LSTM work to capture the sequence information in the segments.

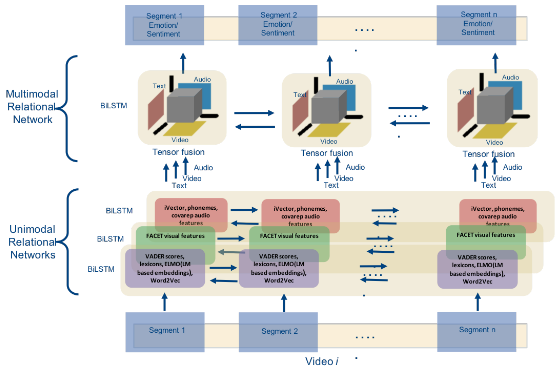


Figure 2: Relational Tensor Network

The other more feasible approach is to apply tensor fusion across modality features of each segment and then model the sequential interactions between the segments of the video using an LSTM Network. We depict our network in Figure 2.

This approach allows generation of contextually rich features that learn their weights not only from the current rich multimodal interactions but also leveraging previous interactions in the process. For example, interactions amongst audio and text features together can have a multiplicative effect to recognize certain kinds of emotion better (for example, high arousal negative words multiplied together can show stronger bias for the angry emotion). Also, these interactions persist across the segments and can help generate more meaningful recognizer of multimodal interactions. The intuitive explanation of this network is that it captures the long term multiplicative effects of interactions across segments for unimodal, bimodal and trimodal features. Neither the TFN model or the contextual model alone can effectively capture these interactions in principle.

	Binary		7-class		Regression
Baseline	Acc	F1	Acc	F1	MAE
$SVM\ multimodal$	60.4	0.61	23.5	0.27	1.38
$LSTM\_uni\_audio$	58	0.52	41	0.37	0.73
$LSTM\_uni\_video$	57.9	0.51	45.9	0.40	0.68
$LSTM\_uni\_text$	64.2	0.60	45.8	0.43	0.618
$LSTM\_earlyfusion$	65.2	0.62	46.6	0.44	0.60
TFN	66	0.62	47.9	0.43	0.58
RTN	**66.8**	**0.63**	**49.17**	**0.45**	**0.58**

Table 1: Sentiment Analysis Model Results

5 Experiments

We present multiple sets of experiments in order to evaluate the different models, impact of textual and audio features on sentiment and emotion prediction. Our training data consists of CMU-MOSEI training set where we do a 90/10 split for validation and early stopping experiments. All our results in this paper are reported on the CMU-MOSEI validation set[1].

5.1 Architecture comparisons

Table 1 and Table 2 show the performance of the various models on sentiment and emotion classification. We have used three LSTM based unimodal baselines, each for audio, video and text modalities. From the table, we see that unimodal-text network outperforms both audio and video modalities for sentiment. Unimodal-text also outperforms SVM multimodal for sentiment analysis, which is an SVM model trained on concatenated features from all the three modalities. The early fusion network is an LSTM based network(an extension of the unimodal networks),that takes in concatenated features from the three modalities. This LSTM model outperforms the SVM multimodal baseline by almost 5% binary class accuracy scores for sentiment analysis. All of these LSTM based networks outperform SVM by a huge margin in the 7-class classification scores and MAE for sentiment analysis. The TFN network with rich set of textual features slightly outperforms the simple concatenation technique(early fusion model) for sentiment and emotion recognition. The model with the best performance is the Relational Tensor Network model for both sentiment and emotion recognition that considers the neighboring tensor fusion networks for a given segment.

5.2 Ablation study

Table 3 shows the detailed ablation study of the various text features that we have used in our models. We added word based features using lexicons and language model based ELMo embeddings and utterance level sentiment scores using VADER scores. As the table shows, adding lexicons result in a slight drop in performance of the scores. The lexicons we've used are extremely sparse compared to the vocabulary space of Word Vectors. Also we've simplistically concatenated the binary scores for Positive and and Negative category words to the same embeddings space as for the word vectors. Majority of these values remain 0 after the operation. We are exploring other ways to leverage the lexicon embeddings to allow a larger contribution of these signals to the classification process. Addition of the ELMo embeddings improves the performance as compared to using word embeddings alone. Addition of ELMo embeddings and segment level sentiment scores using Vader gives the best performance for binary, 7-class and MAE scores, as compared to adding individual features, or a combination of features. As described in ELMo work, adding the layers at different positions of the network helps to abstract various naturally occurring syntactic and semantic information about the words. For the audio modality, we presented two additional feature-sets in the previous section, i-vector features and phoneme level features alongwith COVAREP features. Based on our experiments, we observed that the performance of the Emotion recognition RTN model with all these features were similar but improved slightly for 'Happy' emotion compared to the RTN model without the additional audio features.

6 Conclusion

In this paper we present a novel model called Relational Tensor Network for multimodal Affect Recognition that takes into account the context of a segment in a video based on the relations and interactions with its neighboring segments within the video. We meticulously add various feature set on the word level, that involves language model based embeddings and segment level sentiment features. Our model shows the best performance as compared to the state of the art techniques for sentiment and emotion recognition on the CMU-MOSEI dataset.

[1] The test set was not released at the time of writing.

	Anger	*Disgust*	*Fear*	*Happy*	*Sad*	*Surprise*
SVM multimodal	0.358	0.19	0.21	1.167	0.33	0.171
LSTM_uni_audio	0.17	0.079	0.09	0.475	0.20	0.073
LSTM_uni_text	0.16	0.08	0.086	0.485	0.195	0.068
LSTM_uni_video	0.148	0.08	0.10	0.42	0.208	0.076
LSTM_earlyfusion	0.148	0.078	0.09	0.428	0.19	0.073
TFN	0.147	0.07	0.089	0.466	0.1766	0.074
RTN	**0.137**	**0.065**	**0.072**	**0.422**	**0.176**	**0.059**

Table 2: Emotion Recognition Model Results - MAE scores

	Binary		7-class		Regression
Baseline	Acc	F1	Acc	F1	MAE
Embedding only	64.6	0.60	48.17	0.43	0.595
Emb + Lex	62.8	0.57	45.6	0.41	0.61
Emb + Vader	64.2	0.59	45.4	0.42	0.61
Emb + ELMO	65.5	0.61	47.5	0.44	0.589
Emb + Lex + Vader	64.2	0.59	47.5	0.44	0.59
Emb + ELMO + Lex	64.6	0.58	48.7	0.45	0.576
Emb + ELMO + Vader	**66.4**	**0.63**	**48.9**	**0.44**	**0.577**
All features	66	0.62	47.9	0.43	0.58

Table 3: Text Ablation Study

References

T. Baltruaitis, 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)*, pages 1–10.

Dmitri Bitouk, Ragini Verma, and Ani Nenkova. 2010. Class-level spectral features for emotion recognition. *Speech communication*, 52(7-8):613–625.

G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. 2014. Covarep x2014; a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

Abhinav Dhall, Roland Goecke, Jyoti Joshi, Michael Wagner, and Tom Gedeon. 2013. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM.

P. EKMAN. 2002. Facial action coding system (facs). *A Human Face*.

Paul Ekman. 2009. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. WW Norton & Company.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*. The AAAI Press.

Lakshmish Kaushik, Abhijeet Sangwan, and John HL Hansen. 2013. Sentiment extraction from natural audio streams. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8485–8489. IEEE.

Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. 2004. Emotion recognition based on phoneme classes. In *Eighth International Conference on Spoken Language Processing*.

Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren. 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1695–1699. IEEE.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors.

Michael Neumann and Ngoc Thang Vu. 2017. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matthew Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Robert Plutchik. 2001. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *ACL*.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(03):715–734.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, EPFL-CONF-192584. IEEE Signal Processing Society.

Klaus R Scherer. 1984. Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality & Social Psychology*.

Björn W Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K Burgoon, Alice Baird, Aaron C Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Interspeech*, pages 2001–2005.

Bonggun Shin, Timothy Lee, and Jinho D. Choi. 2017. Lexicon integrated cnn models with attention for sentiment analysis. In *WASSA@EMNLP*.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Fei Tao and Gang Liu. 2017. Advanced lstm: A study about better time dependency modeling in emotion recognition. *arXiv preprint arXiv:1710.10197*.

Fei Tao, Gang Liu, and Qingen Zhao. 2018. An ensemble framework of voice-based emotion recognition system for films and tv programs. *arXiv preprint arXiv:1803.01122*.

Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

Rui Xia and Yang Liu. 2016. Dbn-ivector framework for acoustic emotion recognition. In *INTERSPEECH*, pages 480–484.

A Zadeh, PP Liang, S Poria, P Vij, E Cambria, and LP Morency. 2018a. Multi-attention recurrent network for human communication comprehension. In *AAAI*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *CoRR*, abs/1707.07250.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927*.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018c. Human multimodal language in the wild: A novel dataset and interpretable dynamic fusion model. *Association for Computational Linguistics*.

Amir Zadeh, Paul Pu Liang, Jon Vanbriesen, Soujanya Poria, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. 2018d. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Association for Computational Linguistics (ACL)*.