Sentimetalchemy (STM): A Multi-modal Approach to Sentiment Analysis and Emotion Recognition

Tongxi Liu†, Yutong Li†, Lexie Wang†, Kexin Gao†, Gina-Anne Levow, Haotian Zhu
Department of Linguistics
University of Washington
{ltxom, lyt826, lexwang, kexing66, levow, haz060}@uw.edu

Abstract

Multimodal sentiment analysis (MSA) and emotion recognition in conversation (ERC) tasks aim to identify emotions and opinions in natural language processing (NLP). In this paper, we explore different approaches to MSA and ERC, implementing uni-modal classifiers and fusing multiple modalities. The models are evaluated on the CMU-MOSI and CMU-MOSEI datasets. Results show that deep learning models outperform the SVR-based baseline, and multimodal fusion improves performance. The addition of attention mechanisms further enhances the models' capabilities. The adapted model achieves competitive results on sentiment analysis and emotion detection tasks. This study provides insights into the model architecture and performance evaluation for MSA and ERC.

1 Introduction

Multimodal sentiment analysis (MSA) and emotion recognition in conversation (ERC), the subfields of natural language processing (NLP), aim to automatically identify and categorize emotions, attitudes, and opinions conveyed in textual data. Traditionally, sentiment analysis (SA) has primarily relied on textual information to infer sentiments. However, with the rise of social media, image sharing and multimedia content, including visual and acoustic modalities, have proven valuable in capturing a more holistic understanding of sentiment.

MSA, an emerging research area, integrates multiple modalities such as text, images, audio, and video to uncover human sentiment's rich and nuanced aspects. By combining these diverse sources of information, MSA techniques strive to achieve a more accurate and comprehensive representation of human emotions and opinions expressed on online platforms.

In this project, we discover different approaches to MSA and ERC. We implement uni-modal classifiers using statistic-based machine learning models and neural-based models. We also discover a way to fuse different text, video, and audio modalities to robust the classifier. We analyze the performance of the current state-of-art model UniMSE (Hu et al., 2022) and discover new strategies for this task. Lastly, we perform an adaptation task on MSA and ERC on a larger dataset.

2 Task description

2.1 Dataset

In order to construct sentiment classifiers utilizing a fusion of different modalities, we have selected two multimodal datasets that suit our tasks.

The CMU-MOSI dataset is a relatively small collection comprising 2199 opinion video clips, specifically curated for MSA task (Zadeh et al., 2016). The dataset offers annotations of sentiment ranging from -3 to 3.

Another dataset, CMU-MOSEI, stands as the largest dataset available to date for MSA and ERC (Bagher Zadeh et al., 2018). CMU-MOSEI encompasses over 23,500 sentence utterance videos spoken by a diverse set of over 1000 YouTube speakers. The sentence utterances are randomly sampled from various topics and monologue videos. The videos are transcribed and adequately punctuated, and the dataset exhibits gender balance throughout.

2.2 Main Tasks

We first construct three uni-modal classifiers on textual features to tackle the SA task. Then, we leverage the fusion of three modalities, including text, video, and audio, to enhance the models' robustness.

As our baseline strategy, we develop a Support Vector Regression (SVR) model (Joachims, 2005) based on the textual data extracted from the

†The first four authors equally contributed to this work.

CMU-MOSI dataset. In addition, we implement two deep learning models for comparison: a fully-connected neural network model (NN) (Rumelhart et al., 1986), and a Long Short-Term Memory model (LSTM) (Hochreiter and Schmidhuber, 1997), both trained on the textual features solely. Furthermore, to achieve further enhancements, we construct a multimodal classifier (Fusion) by incorporating features extracted from the text, video, and audio modalities of the CMU-MOSI dataset.

2.3 Adaptation Tasks

After achieving the goals in the main task, we take a significant stride by leveraging the larger dataset, CMU-MOSEI, which is an enhanced version of MOSI. Unlike MOSI, CMU-MOSEI provides annotations for both sentiment intensity and emotion labels. To enhance our multimodal model, we incorporate attention layers (Vaswani et al., 2017), enabling improved handling of the fusion of different modalities. We train this upgraded model separately on CMU-MOSEI, addressing two distinct tasks: MSA and ERC.

Building upon the insights and analyses derived from the previous multimodal model, the introduction of attention mechanisms in our new model may facilitate a more effective combination of information from the text, video, and audio modalities.

To evaluate our models' performance, we compare our models' results against those of state-of-the-art models that have demonstrated high performance on the CMU-MOSEI dataset (Hu et al., 2022). Through this comparison, we anticipate gaining insights into the strengths and limitations of our model architecture. Consequently, this analysis will guide us in identifying potential improvements in data preprocessing techniques, architecture design, and parameter tuning.

2.4 Evaluation Metrics

For the main tasks on MOSI and the adaptation task on MOSEI, we follow the evaluation methods in previous works (Han et al., 2021; Hu et al., 2022), using mean absolute error (MAE), Pearson correlation (Corr), seven-class classification accuracy (ACC-7), binary classification accuracy (ACC-2) and F1 score as performance evaluation metrics. We will also analyze model limitations, ethical risks, and future work of our study.

For the adaptation task, we calculate the accuracy of the model's prediction on each of the six labels. For the emotion annotation of MOSEI, one data point can have multiple labels, but our models can only predict the single-class result. Therefore, we take a prediction as correct if the predicted label is one of the gold labels.

3 System Overview

Our system consists of three primary components. In the initial phase, we utilize CMU-Multimodal SDK to load the datasets. As the datasets encompass text, video, and acoustic feature data with varying frequencies, we designate text as the pivot modality and align the acoustic and visual features accordingly. Moving forward, we partition the datasets into train, development, and test sets using the default portion suggested by the SDK. The second component encompasses baseline and multimodal models that take vector-based representations as input and generate predictions as output. Lastly, we assess the performance of the models and visualize any errors encountered during the evaluation process.

4 Approach

4.1 Uni-modal

We establish three baseline sentiment classifiers, employing Support Vector Regression (SVR), Fully Connected Neural Networks (NN), and LSTM network, all trained on the text data from the CMU-MOSI dataset.

To extract text feature vectors from the CMU-MOSI dataset, we align them with the corresponding labels. The dimensionality of the word embeddings is reduced from $n \times 300$ to 1×300 by taking the average. For Neural Networks, we linearly transform the output labels from the original range of [-3, 3] to [0, 1] to enable the application of the sigmoid function in the output layer. The dataset is split into train (58%), test (10%), and development (32%) subsets based on the GOLD metrics provided by CMU-MOSI.

For the SVR model, we conduct hyperparameter tuning on the development split, exploring various kernel options such as 'linear,' 'poly,' 'rbf,' and 'sigmoid,' along with tuning the kernel coefficient (gamma), epsilon, and squared l2 penalty (C) using grid search. The coefficient of determination of the predictions is then calculated on the test set. Additionally, we employ grid search to tune hyperparameters such as batch size, number of epochs, number of layers, layer size, and activation function for the Feed-forward Neural Networks.

4.2 Multi-modal

We train a multimodal classifier called STM-Fusion to capture a broader context of information beyond the baseline uni-modal models.

In addition to extracting text feature vectors from the CMU-MOSI dataset, we also extract visual and acoustic feature vectors and align them with the corresponding text data. The size of each text feature vector is $n \times 300$, where n represents the length of the instance and varies within the dataset. By utilizing the alignment function provided by the CMU-Multimodal SDK, the dimensions of the visual and acoustic feature vectors are transformed to $n \times 47$ and $n \times 74$, respectively. We combine these three types of feature vectors and split the data into train, development, and test datasets.

The architecture of our fusion model is illustrated in Figure 1. After preparing the data, we pass the text, visual, and acoustic features through unified feature extractor layers. Subsequently, we perform late modality fusion and feed the resulting representation into three LSTMs and two fully connected layers.

Similar to the approach mentioned in section 4.1, we utilize the development set to fine-tune hyperparameters, including batch size, hidden layer size, dropout rate, weight decay, number of layers, and epochs.

We trained another multimodal model using the attention mechanism on the CMU-MOSEI dataset. Like the fusion model, it aligns modalities and passes them into three individual LSTM layers, as illustrated in Figure 2. The critical differences between these two multimodal models are the source of visual feature embeddings and the fusion processes. The fusion model uses an $n \times 47$ Emotient FACET (iMotions, 2017) containing a set of six basic facial emotions. In contrast, the attention model uses MultiComp OpenFace (Baltrusaitis et al., 2016), which provides an $n \times 713$ visual feature representation with more visual semantic information (e.g., facial landmarks, facial shape pa-

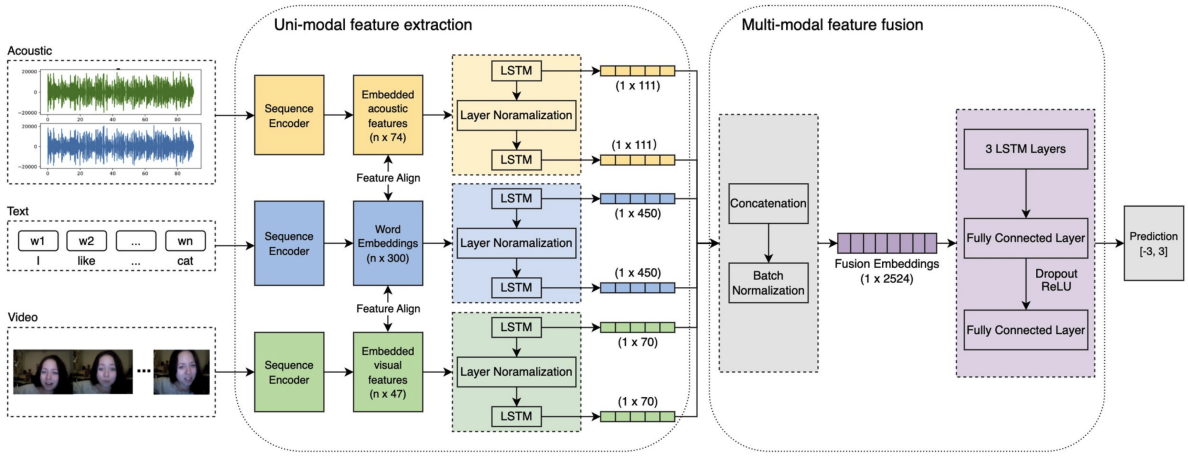

Figure 1: Overview of Fusion Model Architecture

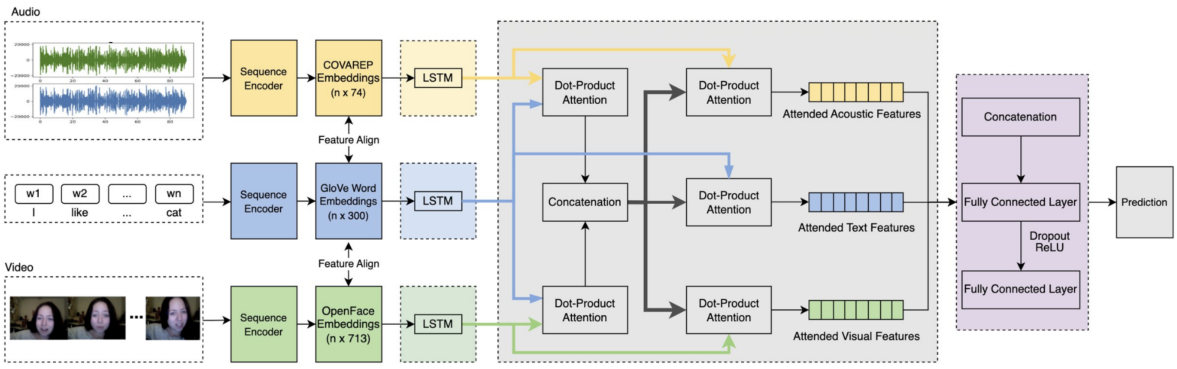

Figure 2: Overview of Model Architecture with Attention Mechanism

rameters, head pose, etcetera). The attention model introduces dot-product attention layers, rather than simply concatenating the output of each modality from the LSTM layers, to learn the attention scores of each modality.

Five dot-product attention layers are applied to fuse the outputs from three modalities. The first two attention layers take the output from the acoustic and textual output and the output from the visual and textual outputs, respectively. We used textual output twice because we found that textual data can convey more information to the MSA and ERC tasks. The result of these two attention layers will be concatenated and input into three more dot-product attention layers. Each layer takes the output from each modality's LSTM layer again and produces attended acoustic, attended textual, and attended visual features. Eventually, we concatenate these attended features into one representation and process it with the same architecture in the fusion model.

5 Results

We set the optimal parameters for each model by grid searching. We choose the Radial Basis Function kernel for the SVR model, C of 200, epsilon of 1, and automatic gamma value. For the fully-connected neural networks, we set up two hidden layers with sizes of 512 and 256, respectively, followed by the ReLU activation function. The sigmoid function is applied to the output layer. For the LSTM model, we implement one LSTM layer (with ReLU activation and dropout of 0.5), followed by two fully connected layers (the former uses ReLU, and the output one uses Sigmoid).

Table 2 shows the performance of each model on the CMU-MOSI test set, compared with the current state-of-art multimodal model, UniMSE (Hu et al., 2022).

Deep learning models (i.e., fully-connected Neural Networks and the LSTM model) significantly outperform the baseline SVR model in ACC-2 for the uni-modal method. In contrast, the SVR model gives better results on ACC-7. LSTM leads to higher ACC-7 compared to SVR and NN.

Comparing the confusion matrix from the binary classification results, the neural network model has more true positive predictions than the SVR model. In contrast, SVR has fewer false positive predicts and has more correct predictions in negative sentiment, which indicates that, in distinguishing negative affect, the SVR model is still competitive compared to the neural network model.

By comparing the training and validation loss over epochs (Figure 3), it also shows that the LSTM architecture significantly improves the over-fitting issue that exists in the vanilla NN architecture, which indicates that STM-LSTM is more robust to generalizing unseen data.

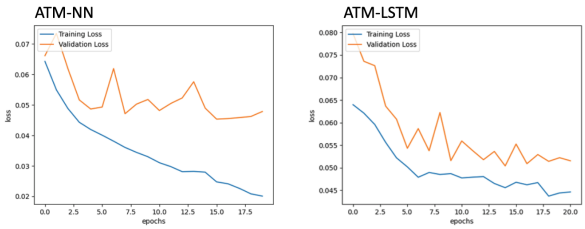

Figure 3: Loss vs. Epochs of STM-NN and STM-LSTM

The multimodal model Fusion adds two other modalities (i.e., video and audio). We expect the results to outperform the uni-modal methods, assuming two extra feature resources could lead to more accurate classification. The ACC-7 of the fusion model beats the one from LSTM, whereas its ACC-2 only slightly outperforms the SVR baseline.

Though there is still a gap between our results

Method	**Model**	**MAE** ↓	**Corr** ↑	**ACC-7** ↑	**ACC-2** ↑	**F1** ↑
Uni-modal	STM-SVR	1.60	0.39	20.52	70.74	72.88
Uni-modal	STM-SVR*	0.70(0.67)	0.34(0.29)	40.1(42.5)	62.5(59.3)	77.4(75.7)
	STM-NN	**1.02**	**0.43**	20.08	73.80	75.63
	STM-LSTM	1.09	0.37	21.40	**74.24**	**75.69**
Multi-modal	STM-Fusion	1.08	0.31	22.16	70.99	71.73
	STM-Attention*	**0.68**(**0.65**)	0.35(0.32)	**44.5**(**46.6**)	62.8(60.1)	**78.9**(**77.9**)
	*UniMSE**	*0.69*	*0.81*	*48.68*	*86.90*	*86.42*

Table 1: Results on MSA. * denotes that the model is trained on MOSEI (the others are trained on MOSI). Contents in parentheses denote the accuracy of the dev sets. Contents in italics denote the current SOTA model. Contents in bold denote the best performance from our models.

and the SOTA model of UniMSE, this is already a relatively good performance, given that our models are trained on a smaller dataset.

The attention model trained on CMU-MOSEI outperforms the STM-SVR baseline model across multiple evaluation metrics, including F1 score, R-squared, ACC-2, and ACC-7. Conversely, our STM-Fusion model only surpasses the SVR baseline model in ACC-2 and ACC-7 on the CMU-MOSI dataset. The attention mechanism is particularly efficient in predicting the seven class sentiments (ACC-7 = 44.5).

In ERC, we employed the same attention-based model architecture used in MSA. The dataset labels consist of a vector size of 6, representing emotions including happiness, sadness, anger, fear, disgust, and surprise. Each vector element is annotated on a scale of [0, 3], indicating the degree of presence of that emotion. Our model prediction is accurate if the predicted emotion has an annotation value greater than 0. As shown in Table 2, the attention model (ACC = 62.9) performs slightly better than the SVR baseline model (ACC = 62.6). In addition, we conduct experiments using single modalities, including text, visual, and acoustic. As anticipated, the attention model's performance surpasses unimodal LSTM models.

Model	**Modality**	**ACC**
SVR	Text	62.6 (63.1)
LSTM	Text	61.8 (62.9)
LSTM	Acoustic	62.7 (62.3)
LSTM	Visual	62.8 (62.4)
Attention	T+A+V	**62.9** (**63.4**)

Table 2: Results on ERC. Contents in parentheses denote the accuracy of the dev sets.

6 Discussion

To briefly recapitulate the improvements we made for our adaptation task, we added the attention mechanism to our previous multimodal STM-Fusion model to train a multimodal attention-based model. Also, we trained our model on the more extensive and diverse CMU-MOSEI dataset instead of the CMU-MOSI dataset we used previously. Additionally, since the CMU-MOSEI dataset also includes emotion labels as a y variable, our multimodal attention-based model can be trained to perform two tasks separately: MSA and ERC.

Concerning the MSA task, our multimodal attention-based model outperforms our SVR baseline on all metrics, including $F1$, MAE, ACC-2, ACC-7, and R^2. The attention-based model (ACC-2 $= 62.8$) performs slightly better than the SVR baseline (ACC-2 $= 62.5$) on binary classification, whereas the improvement is more evident on seven-class classification task (ACC-$7_{Attention} = 44.5$, ACC-$7_{SVR} = 40.1$). The attention-based model ($F1 = 78.9$) also outperforms the SVR baseline ($F1 = 77.4$) on the $F1$ score.

It is important to emphasize that the SVR baseline is trained on texts only, whereas the attention-based model is trained on three modalities of data. The performance improvement is likely since the multimodal attention-based model leverages multiple sources of information, including text, audio, and video. By incorporating diverse modalities, the attention-based model can capture different aspects and perspectives of the data. This integration of complementary information can lead to a more comprehensive understanding of the underlying patterns and structures in the data.

For the ERC, we trained one SVR baseline on texts only, and we also trained three uni-modal LSTM models on each of the three individual modalities. We trained our attention-based model to perform ERC as well. The multi-modal attention-based model outperforms all other models, including SVR$_{text}$, LSTM$_{text}$, LSTM$_{audio}$, and LSTM$_{video}$. The attention-based model has the highest accuracy score on the evaluation set (ACC=63.4) and also the development set (ACC=62.9), which are slightly better than the scores achieved by other models.

Given that the attention-based model performs only slightly better than the SVR baseline and the uni-modal LSTM models, it is very likely that the three modalities only provide a limited extent of complementary information but lack additional complementary information, which explains why the multimodal attention-based model does not have a significant advantage over the uni-modal ones.

Other problems limit the performance of our models. The first problem is an imbalance in the distribution of emotion labels across modalities, which can limit the model's ability to learn from multiple modalities of information effectively. In the CMU-MOSEI dataset, 62.3% of emotion labels is "happy," 16.4% is "sad," 11.7% is "angry," and the rest is distributed among "disgust," "sur-

prise," and "fear." The main problem with having an imbalanced dataset is that models trained on imbalanced datasets tend to favor the majority class, as the model saw most examples from the majority class. Thus, the model struggles to make predictions of minority classes, which leads to biased model predictions and poor generalization.

Our model's predictions of emotion reflect this problem. Our attention-based model could correctly predict the majority class ("happy") for 94.3% of the time. For the second most frequent class ("sad"), the model only made 26.4% correct predictions. For the third most frequent class ("angry"), the model made a brutal 0.8% of correct predictions. The model never generated any predictions for the remaining minority classes ("disgust," "surprise," and "fear"). These results illustrate how our model struggles with minority class predictions due to the imbalance in the dataset.

To mitigate this problem, we tried to sample from the dataset (e.g., include fewer samples of "happy") to create a more balanced, uniformly distributed dataset. However, this could have improved the model's performance. Eventually, we decided to stick to the entire dataset.

Besides the limitations discussed above, our adaptation task has had many successes. To begin with, we re-implemented our multimodal model using TensorFlow (instead of PyTorch), which provides a better API to implement the attention mechanism. Additionally, to mitigate the problem of limited computing resources and extensive training data, we first experimented on 10% of the dataset. Then we trained the entire dataset, which proved helpful, as training complicated models on large datasets can be computationally expensive and time-consuming. By starting with 10% of the data, we quickly prototyped and experimented with different models or techniques without requiring extensive computational resources.

Most importantly, we designed an optimized training approach to train more than $30\,\mathrm{GiB}$ of embeddings. We tuned parameters like the number of epochs and batch size to find numbers that balance model performance and computing resources required. Finally, we overcame the bottleneck problem we encountered in the main task by integrating the attention mechanism, which led to improvements in model performance, as our adaptation model outperforms the baseline model in all aspects. In contrast, our main task model only outperforms the baseline model on several metrics.

7 Limitations and ethical considerations

Despite designing several model architectures and gradually improving our results, our project has two main limitations:

- We encountered challenges in balancing the label distribution for ERC, significantly impacting our predictions' accuracy. Dealing with imbalanced datasets is a common issue in machine learning, and we recognize the need to enhance our knowledge and experience in addressing this problem effectively.

- Effectively combining and leveraging information from multiple modalities remains a significant challenge in multimodal machine learning models. Although our attention model showed better performance in addressing this issue than the previous fusion model, there is still ample room for improvement.

Furthermore, it is crucial to prioritize ethical considerations in machine learning projects to ensure fairness, transparency, and responsible deployment. Considering energy consumption during model training is essential for promoting sustainable and environmentally conscious practices. Additionally, conducting a thorough scrutiny of the dataset being used is essential. Despite being popular datasets in the MSA field, we discovered that MOSI and MOSEI need more detailed data statements (Bender and Friedman, 2018) that describe the demographic and situational information of the speakers. It is highly recommended to critically examine the datasets' limitations in future works.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. pages 1–10.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward

mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

iMotions. 2017. Facial expression analysis.

Thorsten Joachims. 2005. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*, pages 137–142. Springer.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259.