viEMO: Advancing Emotion Recognition in Vietnamese Language through Multimodal Learning

Nguyen Phuc Mac[1,2], **Qui Hung Kieu**[1,3], **Duy Liem Nguyen**[1,4]

[1] Faculty of Information Science and Engineering,
University of Information and Techonology
[2] <22521123@gm.uit.edu.vn> [3] <22520505@gm.uit.edu.vn>
[4] <22520752@gm.uit.edu.vn>

Abstract

This study investigates Multimodal Emotion Recognition (MER) in the Vietnamese context, combining text, speech, and image modalities. We utilize deep learning models and implement fusion methods - product fusion and weighted average fusion. While initial results show lower accuracies compared to the best-performing unimodal models, the analysis reveals the potential of multimodal fusion for improving emotion recognition beyond single-modality approaches. This research provides a foundation for future work on MER in Vietnamese, emphasizing the need for further research on effective fusion techniques and robust feature extraction methods.

Keywords: multimodal emotion recognition, deep learning, product fusion, weighted average fusion, Vietnamese dataset, human-computer interaction, emotion detection.

1 Introduction

Multimodal Emotion Recognition (MER) is a crucial research area with significant potential applications in human-computer interaction, healthcare, and user experience analysis. MER combines information from multiple modalities such as text, audio, and visual data to enhance the accuracy of emotion recognition. Relying solely on a single modality, such as text, can be insufficient to fully capture the true emotional state of the speaker, as textual content alone may not accurately reflect the underlying emotions conveyed through nonverbal cues like tone of voice or facial expressions. Moreover, existing research on MER has primarily focused on languages other than Vietnamese. The unique structure and semantics of the Vietnamese language present specific challenges in emotion recognition, requiring tailored approaches and datasets.

To address these limitations and contribute to the advancement of MER in the Vietnamese context, this study investigates the development of a MER system that integrates information from text, speech, and image modalities. We utilize deep learning models and implement fusion techniques, including product fusion and weighted average fusion, to effectively combine the strengths of each modality. To support this research, a novel Vietnamese trimodal dataset has been created. This dataset features four distinct labels: three individual labels for each modality (text, speech, image) and one unified label determined by observing the combined information across all three modalities. Furthermore, the dataset includes the "neutral" emotion as a classification category for instances where the expressed emotion falls outside the range of the six primary emotions (happiness, sadness, anger, fear, surprise, and disgust). While the initial results may not have met our expectations, this research endeavor provides valuable insights and lessons for future advancements in multimodal emotion recognition, particularly within the Vietnamese context.

2 Fundamental

2.1 Related works

Research on multimodal emotion recognition has garnered significant attention within the scientific community, with substantial advancements in combining information from modalities such as text, audio, and images. This section provides an overview of foundational research, from initial contributions to recent developments, while highlighting the challenges that still persist in the field. A comprehensive survey by Ahmed et al. (2023) systematically analyzes various learning algorithms, fusion techniques, and datasets used in multimodal emotion recognition research, identifying key trends, challenges, and promising research directions. This survey provides valuable insights into the current state-of-the-art, guiding future research efforts in this dynamic field.

One notable study is by Anatoli de Bradké, Mael

Dataset	Year	Modality	Language	Topic	#Samples
CMU-MOSEI	2018	A+V+T	English	Sentiment	23,453
MELD	2019	A+V+T	English	Emotion	13,000
CH-SIMS	2020	A+V+T	Chinese	Sentiment	2,281
CMU-MOSI	2016	A+V+T	English	Sentiment	10,000
IEMOCAP	2008	V	English	Emotion	302

Table 1: Summary of reference datasets. Note: A = Audio, V = Video, T = Text.

Fabien, Raphae Lederman, and Stéphane Reynal (2018-2019). This research utilized text data from daily writings of psychology students, audio from the RAVDESS database, and emotional images from the FER2013 dataset. The approach encompassed data preprocessing, feature extraction, and the application of machine learning and deep learning algorithms for emotion classification. While this research laid an important foundation, the integration of modalities (fusion) was not implemented comprehensively, leading to limitations in optimizing model effectiveness.

Building upon foundational datasets like CMU-MOSEI, MELD, CMU-MOSI, CH-SIMS, and IEMOCAP has significantly advanced multimodal emotion recognition research. CMU-MOSEI and MELD stand out with their comprehensive annotations, facilitating in-depth analysis of emotion and sentiment in multimodal contexts. While smaller in scale, CMU-MOSI has garnered considerable attention for its detailed annotations, particularly within sentiment analysis research. Recognizing the need for culturally sensitive systems, CH-SIMS provides a valuable multilingual resource specifically tailored for Chinese sentiment analysis. IEMOCAP, with its emphasis on speech-based emotion recognition, offers high-quality audio data that has proven invaluable for developing voice-driven emotion models.

In the field of text-based emotion recognition, advancements in natural language processing (NLP) have facilitated the development of increasingly accurate feature extraction methods. Studies like "Emotion Recognition for Vietnamese Social Media Text" (Ho et al., 2019) demonstrate the importance of addressing language-specific challenges, highlighting the need for models that can effectively capture the nuances of emotion expression in different linguistic contexts. This research, focusing on Vietnamese social media text, investigates the unique challenges posed by the Vietnamese language, such as its complex morphology and the prevalence of slang and informal language, and explores effective techniques for capturing these linguistic nuances to accurately classify emotions. For audio, databases like RAVDESS continue to be a popular resource, supporting research on emotions through voice. However, for images, the FER2013 dataset has faced criticism due to low resolution, data imbalance, and issues with label reliability.

Methods for combining information from multiple modalities have also been explored, including weighted average fusion and product fusion. These methods hold promise for significant improvements in multimodal emotion recognition. However, the major challenge lies in optimizing these techniques to ensure accuracy and efficiency in real-time applications.

Despite notable progress, the field still faces numerous opportunities and challenges in developing more complete systems that meet the demands of complex emotion analysis in real-world scenarios.

2.2 Methodology

In this study, we employ feature extraction methods to recognize emotions from audio, facial expressions, and text. For audio emotion recognition, we utilize Mel-Frequency Cepstral Coefficients (MFCC) features to capture the spectral characteristics of audio signals. These features are extracted using the librosa library and compiled into multidimensional vectors that serve as input for deep learning models. The "Toronto Emotional Speech Set" dataset is employed for training. In facial emotion recognition, we apply convolutional neural networks (CNNs) with Conv2D, MaxPooling, and Flatten layers to extract facial features. The data is subsequently converted into 1D vectors for emotion prediction through fully connected layers, with dropout layers to prevent overfitting. The predict_emotion function is used to process images and predict emotions from the model. For text-based emotion recognition, we utilize the UIT-

VSMEC dataset from social media language, followed by normalization to regular language using ViLexNorm and BARTpho-syllable. PhoBERT is used to train the model for Vietnamese text emotion recognition. Finally, fusion methods are employed to combine the predicted ratios and arrive at a final conclusion. A comparison with the evaluated labels is then conducted to provide feedback and identify areas for improvement, as well as the strengths and weaknesses of this approach.

3 Dataset

3.1 Dataset creation

In this study, we constructed a multimodal dataset from conversational videos on YouTube, focusing on clips where the speaker's face, voice, and speech could be clearly extracted. The data collection process consisted of three main steps. First, we identified the start and end times of each conversational segment, then extracted the audio from these segments to obtain sound files. Noise reduction and normalization techniques were applied to ensure quality. Next, the speech in the audio was transcribed into text using the Speech Recognition (SR) library and manually checked to improve accuracy. Finally, we extracted images containing the speaker's face, retaining only high-quality frames that clearly displayed close-up facial features.

After collection, we employed the following guidelines for emotion annotation: 1) Prioritize the primary, most dominant emotion, considering facial expressions, vocal cues, and the conversation's context. 2) Ensure objective and consistent interpretations with a focus on inter-annotator reliability. 3) Utilize clear and concise instructions with examples and edge cases. 4) Implement quality control with multiple annotators per clip and inter-annotator agreement checks. 5) Continuously review and refine the guidelines based on feedback. To assess initial annotation reliability, we conducted a validation process on 100 randomly selected data points with three independent annotators. The results, evaluated using Cohen's kappa coefficient, achieved a value of 0.770, indicating a fairly good level of agreement and demonstrating high consistency.

Cohen's kappa is commonly used to measure agreement between two raters. However, when more than two raters are involved, an extended version, called Fleiss' kappa, can be employed to evaluate agreement among multiple raters.

Attribute Name	Description
start_time	Start time of video segment
end_time	End time of video segment
Emotion_Text	Emotion label for text
Text	Spoken words in the video segment
Emotion_Audio	Emotion label for audio
Audio	Path to the audio file of the video segment
Emotion_Image	Emotion label for image
Image	Path to the image file of the video segment

Table 2: Description of dataset attributes.

Fleiss' Kappa is an extension of Cohen's kappa that measures the degree of agreement among multiple raters when categorizing items into mutually exclusive categories. The formula for calculating Fleiss' Kappa is defined as follows:

$$\kappa = \frac{\bar{P}_o - \bar{P}_e}{1 - \bar{P}_e}$$

Included:

- P_o: The observed agreement ratio (Observed Agreement).
- P_e: The expected agreement ratio (Expected Agreement) if labels are assigned randomly.

The Kappa coefficient measures the level of agreement between different raters. The value of Kappa ranges from -1 to 1, where:

- $\kappa = 1$: Complete agreement.
- $\kappa = 0$: Agreement occurs entirely by chance.
- $\kappa < 0$: The observed agreement is lower than the agreement expected by chance.

3.2 Training Data

3.2.1 TESS Dataset

TESS (Toronto Emotional Speech Set) is an audio dataset designed to support research on emotion recognition through speech. The dataset includes recordings of two female actors, one aged 26 and the other aged 64, with the goal of studying emotional differences across age groups. Each recording simulates seven different emotions: happiness, sadness, surprise, fear, anger, disgust, and neutral. The content of the recordings is based on 200

Figure 1: Examples of Multimodal Data from the Dataset.

meaningful sentences from the Harvard Sentences, a collection commonly used in speech processing research. The dataset provides audio files in .wav format, optimized for clear emotion transmission. With this design, TESS becomes a valuable resource for developing and evaluating emotion recognition systems through speech, such as virtual assistants or psychological care applications.

3.2.2 Face emotion image dataset collected from GitHub

For images, we utilized data from Chirag Joshi's GitHub repository at the Face Detection Model link. This dataset contains 35,887 entries, comprising images that showcase a wide range of facial emotional states, including happiness, sadness, anger, and surprise, divided into two sets: train and test. With its richness in quantity and diversity in characteristics, this dataset serves as a vital visual resource for analyzing and recognizing emotions through facial expressions. Furthermore, using open-source data from platforms like GitHub ensures scalability and flexibility throughout the research process.

3.2.3 UIT-VSMEC Dataset

UIT-VSMEC (Vietnamese Students' Mental Emotion Corpus) is a Vietnamese dataset developed to support research on emotion recognition in text. This dataset contains 6,927 sentences or passages, each labeled according to six common emotion categories: happiness, sadness, surprise, fear, anger, and neutral. The dataset's content was collected and annotated by experts, focusing on feedback, comments, and interactions related to student life. UIT-VSMEC is designed for studying and developing natural language processing (NLP) models specifically for Vietnamese, particularly in areas such as chatbots, user feedback analysis, or automatic emotion recognition. It is a pioneering dataset in advancing research on text emotion recognition in Vietnam, while also unlocking significant potential for practical applications.

4 Evaluation

This report uses the Jupyter Notebook environment on the Windows 11 operating system, with a Ryzen 7 processor, 16GB of RAM, and Python 3.11.12 as the programming language version. It also utilizes essential libraries such as matplotlib, seaborn, numpy, pandas, and scikit-learn.

4.1 Speech Emotion Recognition

Feature extraction is a critical step in speech emotion recognition, converting raw audio signals into meaningful representations that machine learning algorithms can process effectively. In this study, we extract audio features by converting the signals into Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs capture the spectral characteristics of the audio signal, representing how humans perceive sound on the Mel scale. MFCCs are widely used in audio processing tasks due to their ability to encompass both low-level and high-level acoustic information. These features are extracted using the librosa library, where the raw audio signal is preprocessed to standardize the sampling frequency before feature computation. The extracted features are then combined into a feature vector, encompassing various aspects of the sound. This multidimensional feature vector is used as input for machine learning models for classification, enabling

Modality	Angry	Disgust	Fear	Happy	Neutral	Surprise	Sad
Emotion_Text	262	94	52	173	216	136	150
Emotion_Audio	233	172	73	77	233	107	188
Emotion_Image	219	82	60	203	159	102	257

Table 3: Distribution of emotion labels across modalities.

emotion recognition embedded in the audio. Subsequently, we use the 'Toronto Emotional Speech Set' dataset on Kaggle to train the model for emotion recognition based on speech intonation.

4.2 Face Emotion Recognition

In the task of emotion recognition from facial expressions, feature extraction is a crucial step that enables the model to identify essential information from images. Convolutional Neural Networks (CNNs) automatically perform this process through layers such as Convolutional (Conv2D), MaxPooling, and Flatten. Specifically, the Conv2D layers use filters to detect local features in the image, such as edges, corners, and facial details, enabling the recognition of important features such as the eyes, nose, mouth, and facial expressions. The pooling layers (MaxPooling2D) help reduce the size of the features while retaining the most important information, and also reduce the number of parameters, speeding up computation and minimizing the risk of overfitting.

After passing through the convolutional and pooling layers, the data is converted from a 2D matrix to a 1D vector through the Flatten layer, preparing it for the fully connected layers. These layers (Dense) use the extracted features to learn and make predictions about emotions. Finally, a Dropout layer is used to prevent overfitting by randomly removing some connections during training, helping the model generalize better.

The emotion recognition process from facial expressions is performed through the trained model, and the prediction results are made using the predict emotion function. First, this function takes the image path as input and processes the image by converting it to grayscale, then resizing the image to the standard size (48x48 pixels). After processing, the image is normalized, converted to an array, and a batch dimension is added to match the model's input format. The model will predict the probability of each emotion, and the function uses argmax to extract the emotion with the highest probability.

4.3 Text Emotion Recognition

We have reviewed several papers in the field of emotion recognition in Vietnamese. To the best of our knowledge, we have not found any pre-trained models specifically designed for emotion recognition on Vietnamese text. Therefore, we sought a labeled dataset containing emotions associated with each sentence in the Vietnamese language and trained a model to meet this requirement. However, we were unable to find any datasets suitable for academic purposes in standard Vietnamese language for emotion recognition. The most suitable dataset for our task is UIT-VSMEC. This dataset includes 6,927 sentences labeled with emotions based on the six basic emotions proposed by Paul Ekman, along with a Neutral label for emotions that do not belong to the six aforementioned categories or when no emotion is present.

However, this dataset contains social media language, which differs significantly from formal language in terms of usage and structure. Social media language tends to be informal, flexible, and creative, allowing users to employ abbreviations, emoticons (emoji), and slang to express emotions quickly and easily. Sentence structures in social media language may be short, lack complete grammatical components, and sometimes include deliberate spelling mistakes. Social media language is mainly flexible and time-efficient, using abbreviations and skipping some grammar rules for quick communication, while formal language requires adherence to grammar and clear structure, especially in formal contexts.

Therefore, there is a need for a method to convert social media language into formal language to meet the requirements of this paper. We use the ViLexNorm dataset and BARTpho-syllable to normalize the text into formal language, and then apply PhoBERT on the normalized dataset to train the model. After normalization, the UIT-VSMEC dataset still has some imperfections but is now more suitable for our task of emotion recognition in Vietnamese text.

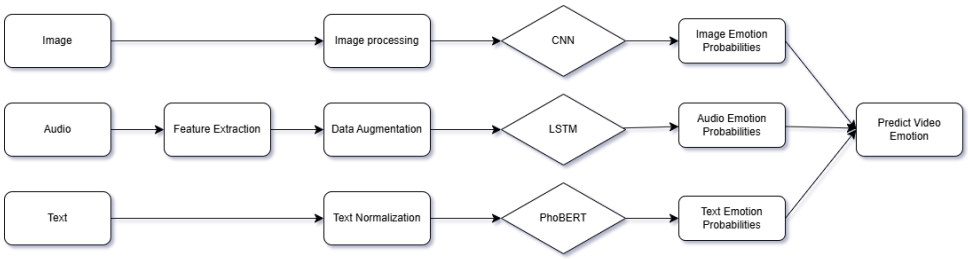

Figure 2: Multimodal Emotion Recognition Pipeline.

5 Multimodel Fusion Method

In a multimodal emotion recognition system, combining information from different modalities plays a crucial role in improving the model's performance and accuracy. We have used two main methods to fuse features from text, audio, and images: Product Fusion and Weighted Average Fusion.

5.1 Product Fusion

Product Fusion is a fusion method in which features from different modalities are multiplied together. The main idea behind this approach is to leverage the interactive relationships between features, allowing them to complement each other and form a stronger unified feature representation.

Specifically, in the task of multimodal emotion recognition, features from text, audio, and images are extracted through separate models (such as natural language processing, audio signal processing, and image processing models). These features are then multiplied across corresponding dimensions to create an integrated feature. This enables the model to take advantage of the interaction between modalities, thereby enhancing its ability to accurately recognize emotions.

However, Product Fusion also presents challenges, such as the risk of increasing the size of the output features or losing information if one modality does not contribute strongly enough to the interaction. Therefore, normalizing and optimizing the features from each modality before applying Product Fusion is an important step to ensure the effectiveness of this method.

5.2 Weighted Average Fusion

Weighted Average Fusion is a flexible and effective fusion method where features from different modalities are combined by calculating a weighted average. The weight of each modality is determined based on its importance in emotion recognition.

In this task, we have found that features from audio and text often carry more decisive information compared to features from images. This is due to the fact that emotions are typically more clearly expressed through voice and textual content. On the other hand, image data, although providing supplementary information through facial expressions, has some limitations. For instance, videos used to collect data are often created by amateur actors, resulting in less diverse facial expressions and sometimes only conveying a single emotion. This can lead to bias if the weight of the image modality is too high in the fusion process.

To address this issue, we have established two specific weight ratios for the modalities:

$(0.5 : 0.35 : 0.15)$ – In this case, the text modality is given the highest priority, followed by audio, and lastly, images.

$(0.35 : 0.5 : 0.15)$ – This case assigns a higher weight to audio, as speech typically carries more natural emotional cues, especially in multilingual data.

These weight ratios are applied flexibly based on the specific scenario and the model's performance on the test dataset. Weighted Average Fusion not only ensures a balanced contribution from all modalities but also helps the model leverage the most relevant information from different data sources without losing the contribution of any modality.

6 Experimental Results

The multimodal emotion recognition system was evaluated based on two main model fusion methods: product fusion and weighted average fusion with two different weight ratios. Both methods were tested on the dataset that our team had collected and labeled with high consensus, aiming to analyze the accuracy of emotion prediction when combining information from multiple modalities such as text, audio, and images. Before the fusion

process, we trained the models for each modality separately. The audio model achieved a remarkably high accuracy of 0.99, demonstrating the strong influence of audio features on emotion recognition. In contrast, the text and image models achieved moderate accuracies of 0.54 and 0.52, respectively, highlighting the challenges associated with accurately interpreting emotions solely based on text and visual cues.

Afterward, we performed fusion of the modalities using the two methods and evaluated the performance on 100 randomly selected samples from our custom-built dataset. To ensure objectivity and practical applicability, we conducted an evaluation on 100 randomly selected samples from the custom-built dataset. Selecting 100 samples helped minimize potential biases caused by the large dataset size and allowed for an in-depth analysis of the performance of each fusion method. Additionally, this approach serves to validate the system's suitability for a smaller yet diverse dataset representing various emotional scenarios.

The results indicate that product fusion provides the most appropriate predictions in alignment with the emotion labels in the dataset, although the similarity between the predictions and the actual labels only reached 20%. While the similarity rate is not very high, product fusion still demonstrates a clear advantage over the other fusion methods, showcasing its ability to effectively utilize information from different modalities. This can be explained by the information combination mechanism of product fusion, where critical information from each modality is computed and processed simultaneously, minimizing the loss of essential data during the fusion process.

In contrast, the weighted average fusion method struggles to maintain consistency and accuracy when integrating information from multiple data sources. This result not only confirms the potential of product fusion but also opens up avenues for further research aimed at improving the similarity between predictions and actual labels. Enhancing the performance of product fusion could be a significant step forward in developing multimodal emotion recognition systems, better meeting practical needs in applications related to user experience, intelligent communication, and advanced emotion analysis.

Method	**Accuracy**
PROD Fusion	0.2
Audio WA Fusion	0.08
Text WA Fusion	0.08

Table 4: Table of comparison ratio between Prediction and Evaluation of Fusion Method.

Emotion	**Precision**	**Recall**	**F1-score**
Angry	0.42	0.16	0.24
Disgust	0.47	0.71	0.56
Fear	0.47	0.55	0.51
Happy	0.70	0.70	0.70
Neutral	0.49	0.29	0.36
Surprise	0.35	0.37	0.36
Sad	0.49	0.56	0.52

Table 5: Results of the text model.

Emotion	**Precision**	**Recall**	**F1-score**
Angry	1.00	0.98	0.99
Disgust	1.00	0.98	0.99
Fear	1.00	1.00	1.00
Happy	0.99	0.98	0.98
Neutral	1.00	1.00	1.00
Surprise	0.95	0.97	0.97
Sad	1.00	1.00	1.00

Table 6: Results of the audio model.

Emotion	**Precision**	**Recall**	**F1-score**
Angry	0.46	0.44	0.45
Disgust	0.90	0.16	0.27
Fear	0.30	0.30	0.30
Happy	0.75	0.83	0.79
Neutral	0.44	0.66	0.53
Surprise	0.44	0.48	0.46
Sad	0.93	0.05	0.09

Table 7: Results of the image model.

7 Conclusion

The research results indicate that the product fusion (prod fusion) method provides the highest accuracy in emotion prediction, aligning well with the pre-labeled emotion tags, compared to other methods in the fusion model. Although the similarity between the predicted results and actual labels only reached 20%, prod fusion still demonstrated superior performance over other techniques. In particular, performance evaluation metrics such as F1 score and accuracy for individual modalities—text,

audio, and images—only achieved above-average levels. This confirms that combining modalities through product fusion effectively leveraged the supplementary information from each modality, significantly improving emotion recognition performance in a multimodal context.

References

Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. *Intelligent Systems with Applications*, 17:200171.

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.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Maël Fabien. 2019. Multimodal emotion recognition. https://github.com/maelfabien/Multimodal-Emotion-Recognition.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2019. Emotion recognition for vietnamese social media text. In *Proceedings of the 17th International Conference on Natural Language Processing (PACLING 2019)*, pages 389–398. Springer.

Chirag Jain. 2023. Face detection model - image dataset. https://github.com/Chiragj2003/Face-detection-model/tree/main/images. Accessed: 2025-01-03.

Pichora-Fuller, M. Kathleen, and Kate Dupuis. 2020. Toronto emotional speech set (tess).

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.