UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Small Language Model Fusion Network for Multimodal Affective Computing

Ankita Gandhi[1], Kinjal Adhvaryu[2#]

1 Research Scholar, CE/IT Engineering, Gujarat Technological University, Ahmedabad, India
ORCID: 0009-0007-0464-1728

2 Professor, Computer Engineering, Gujarat Technological University, Ahmedabad, India.
ORCID: 0000-0002-9270-8369

#Corresponding Author: Kinjal Adhvaryu

Abstract

Conventional affective computing models aim to identify sentiments, emotions, hate speech, fake news from written text while Multimodal affective computing models identify emotions, sentiments, opinions expressed in form of multimodal data like images with captions, memes, videos, audios, emojis, texts, physiological signals, etc. In multimodal setup other modalities like speech, visuals accompany text modality. With the invent of language foundation models like ChatGPT and other small language models like Phi 3 mini, Llama, Gemini, etc., the potential of these models can be used to perform affective computing tasks. This study is steered on one specific multimodal affective computing tasks of sentiment analysis using the potential of small language models. The proposed approach employs two different subnetworks and the results from subnetworks then fused to get a more comprehensive understanding of the associated sentiment. One is a language subnetwork which uses small language model as a base model and the other is audio-visual subnetwork. To validate the proposed framework named small language model fusion network (SLMFN), extensive experiments are performed on two benchmark multimodal datasets, namely CMU-MOSI and CMU-MOSEI. This study offers insights into the practical applications of small language models by fine-tuning it for language specific tasks, which advances sentiment analysis and emotional recognition techniques. Additionally, with the use of quantized small language model making the proposed model more suitable for mobile and edge device-based application.

Keywords: *Affective Computing, Sentiment Analysis, Multimodal Fusion, Small Language Models*

I. Introduction

Affective computing involves the exploration and creation of systems and devices capable of identifying, processing, mimicking and interpreting human emotions. This multidisciplinary domain integrates aspects of computer science, design, human psychology, and cognitive science. The scope of research in affective computing encompasses sentiment analysis, emotion detection, opinion mining, sarcasm detection, humor detection and many more. A key motivation behind this field is to endow machines with emotional intelligence and the capacity to simulate empathy. Such systems aim to understand human emotional states and adjust their behavior accordingly. Multimodal affective computing [1] is a technology which includes modalities such as audio and visual data along with text data. It also includes other modalities like EEG signals, physiological signals, emojis, etc. It can be bimodal, involving diverse combinations of two modalities, or trimodal, which integrates three different modalities. It can also be multi-modal includes different combinations of modalities. The emerging fields of affective computing and sentiment analysis, which leverage on human-computer interaction, information retrieval, and multimodal signal processing for distilling people's sentiments from the growing amount of online social data [2]. Sentiment analysis concentrates on identifying the polarity (positive, negative, or neutral) and the intensity of the sentiments expressed and gives insights into user opinions [21]. Emotion recognition involves identifying an individual's underlying emotional or affective state based on their verbal and non-verbal cues, including facial expressions, body language, and speech [3].

With the rise of language foundation models like GPT 3.5[4], Llama [5], Roberta [6], a new artificial intelligence paradigm has emerged, by simply using general purpose foundation models with prompting to solve problems instead of training a separate machine learning model for each problem [7][13]. But the biggest problem with these models is the size and training time to tune for any specific task. Pretrained Language Models have expanded to hundreds of billions of parameters, like GPT-3

[26], demonstrating exceptional few-shot performance. However, training and utilizing such large models demand immense computational resources, leading to a significant carbon footprint and posing challenges for researchers and practitioners to adopt them. Recently, the emergence of small and medium-sized models, which are highly efficient (1B to 3B parameters), has enabled their application in affective computing and sentiment analysis tasks. These smaller language models are considered much "greener" due to their significantly lower parameter count compared to Large Language Models, earning them the designation of small language models. These models are highly lightweight, including their pre-trained and instruction-tuned variants, which makes them well-suited for edge and mobile devices. Interestingly, they exhibit emergent properties that allow them to solve language specific tasks they were not specifically trained on. However, research on the usage and effectiveness of such models remains relatively limited. In this paper, we introduce a new multimodal fusion technique, termed SLM Fusion Network (SLMFN), which performs end-to-end affective computing tasks. It comprises two sub networks, one is language subnetwork which works on textual modality and another is audio visual subnetwork which works on audio and visual modalities. Videos from social media networks are used as a source to extract text, audio and visual modalities.

In this study, for the language subnetwork, the capabilities of such small language foundation models are being used to perform tasks like sentiment analysis. The phenomenon of emerging capabilities of SLMs was more pronounced with the utilization of fine-tuning techniques [15]. The performance is comparable to classical Natural Language Processing (NLP) models like Bag-of-Words (BoW) [16], Glove [31] and better than fine-tuned LLMs like BERT [17] and RoBERTa[6]. Another challenge encountered was interpreting the outputs from the responses of language model, as it is difficult to generate only a single word which is labelled for sentiment analysis. The other is audio-visual subnetwork which is comprised of custom layers of LSTM and bidirectional LSTM followed by fully connected layers for audio and visual subnets. Audio and visual features are aligned with other and also with text data. The key contributions of this work are as follows:

- We propose the utilization of small language models as a, enabling it to deliver classification outputs that address affective computing challenges, demonstrated through tasks like sentiment analysis.
- We develop a fusion approach that combines language network responses with audio-visual sub-network outputs to achieve enhanced classification results for affective computing tasks.
- We demonstrate the effectiveness of our method on the publicly available CMU-MOSI [19] and CMU-MOSEI [20] datasets, achieving superior performance in terms of quantitative metrics (MSE, F1 Score, Recall, Accuracy) and qualitative comparisons with baseline approaches.

The rest of the paper is structured as follows: Section 2 reviews related work in multimodal affective computing using foundation language models. Section 3 details the methodology, model architecture, and training pipeline. Section 4 presents experimental setup, datasets and testing framework. Followed by a discussion of key findings and limitations in Section 5. Section 6 concludes with potential future directions.

II. Related Work

We focus on related work within the area of foundation models in affective-computing-related tasks in the multimodal domain.[8] explores a fusion of ChatGPT for multimodal Named Entity Recognition (NER). [9] investigates EmoLLMs, open-source instruction following LLM, by fine tuning the LLMs to perform the Various affective computing tasks. [10] explores capabilities of foundation models to predict affect outcomes based on smartphone sensing data of university students. [11] explores ChatGPT's zero-shot ability to perform affective computing tasks using prompting alone. [12] investigates on how to relate explanations back to multimodal and time-dependent data, how to integrate context and inductive biases into explanations using mechanisms such as attention, generative modeling, or graph-based methods, and how to capture intramodal and cross-modal interactions in post hoc explanations. They also show future research directions in terms of explainable affective computing. [13] investigates the performance of ChatGPT in sentiment analysis and aspect extraction. Moreover, developing an effective evaluation approach to assess language models on regression tasks provides valuable insights for future research efforts.

III. METHDOLOGY

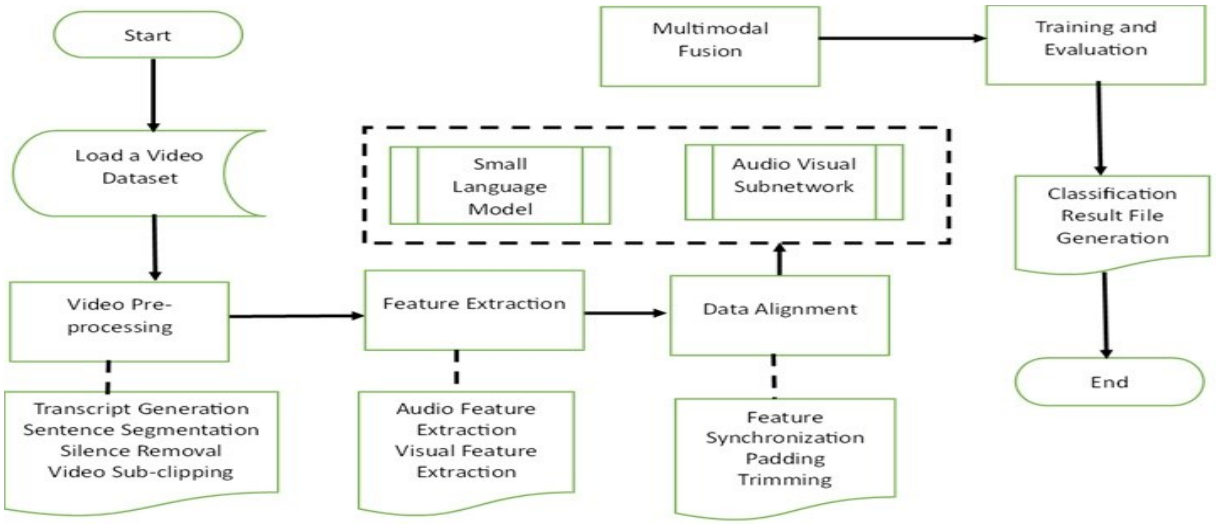

FIG.1 SLMFN Pipeline

Our proposed SLMFN consists of three major parts: 1) Small Language Model to process text modality 2) Audio-Visual subnetwork for processing acoustic and visual modalities 3) Fusion layer to predict final classification. Based on the task of sentiment analysis, the network output adapts to support binary classification, 5-class classification. The input to the SLMFN consists of opinionated videos encompassing three modalities: language, visual, and acoustic. The subsequent three subsections provide a detailed explanation of the SLMFN network.

3.1 Small Language Model

Small language model is built on a transformer-based architecture comprising an encoder-decoder structure. From raw videos, sentence-wise transcripts and visual subclips are generated using Kaldi [22]. For our task, only the decoder part is used, which processes the given text data in sentence form and generates classification results. Decoder architecture enhances contextual understanding. Language Model is fine-tuned to perform sentiment analysis and emotion recognition tasks. Hyperparameter optimization of this architecture is done using Adam [23]. This fine-tuned architecture then generates classification results and performs binary classification and for three-class classification.

3.2 Audio-Visual Subnetwork

The audio-visual subnetwork integrates features from audio and video data for sentiment classification through a synchronized and context-aware process. Audio features, such as MFCCs, chroma, spectral centroid, and Mel spectrogram, are extracted using moivepy and processed using librosa[24] . Visual features like 3D facial features are processed by opencv[25].

3.3 Fusion Layer

Results from the two subnetworks are concatenated to generate final classification results. Binary classification and three-class classification is performed.

IV. Experimental Setup

In this section, the benchmark datasets for the multimodal sentiment analysis are discussed. Afterwards, the testing framework is introduced which includes preprocessing, feature extraction, data alignment and fusion network. Subsequently, how the deep learning models are trained and tuned is discussed. The pipeline of our method is presented in Figure 1.

4.1 Datasets

We present here the adopted datasets for the multimodal sentiment analysis problem. A summary of their statistics is in Table 1.

Name	Modalities	No. of videos	Source	No. of speakers	Language	Topics Covered	Train	Validation	Test
CMU-MOSI	Audio, Visual, Text	93	YouTube	89 41-female 48-male	English	Topics indexed by #vlog	1,284	229	686

CMU-MOSEI	Audio, Visual, Text	3228	YouTube	1000	English	250 Reviews Debate Consulting	16,326	1,871	4,659

Table 1 CMU-MOSI and CMU-MOSEI Dataset Statistics

4.1.1. MOSI Dataset

We make use of the CMU- MOSI (Multimodal Opinion Sentiment Intensity dataset, created by [19], for sentiment analysis tasks. It is one of the pioneering opinion-level sentiment intensity dataset for multimodal sentiment analysis. It includes diverse features such as multimodal observations, transcripts of spoken words, and visual gestures, along with automatically extracted audio and visual characteristics from review videos. A unique aspect of MOSI is its subjectivity segmentation at the opinion level and sentiment intensity annotations. MOSI is composed of 93 YouTube vlog videos, featuring 89 speakers—41 females and 48 males. It is the first dataset of its kind to include opinion-level subjectivity and sentiment intensity annotations, MOSI has set a benchmark for multimodal sentiment analysis research. It has 2199 uttered sentences spoken in the English language and annotated with strongly positive, positive, neutral, negative and strongly negative annotations. The sentiment distribution is approximately 14% strongly negative, 17% negative, 15% neutral, 17% positive, and 13% strongly positive. We have used the original dataset as it is and the training, validation and test portion is as shown in Table 1.

4.1.2 MOSEI Dataset

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [20] dataset marks a substantial step forward in multimodal sentiment analysis and emotion recognition. It is the largest dataset featuring sentence-level sentiment and emotion annotations. This dataset includes over 65 hours of annotated video, featuring 1,000 speakers and 250 unique subjects, with 3,228 videos sourced from YouTube. Each video segment is manually transcribed at the phoneme level and synchronized with audio. The alignment across textual, acoustic, and visual modalities provides a solid basis for thorough analysis. In total, CMU-MOSEI comprises 22,856 sentences from 3,228 YouTube videos, covering diverse topics such as product and service reviews (16.2%), debates on various issues (2.9%), and consulting discussions (2.9%). It stands out as one of the most versatile datasets due to its broad topic coverage. We utilized the dataset in its original form, with training, validation, and test splits detailed in Table 1.

4.2 Testing Framework

For the experimental setup, we implemented a detailed end-to-end workflow for multimodal sentiment analysis. The process began with preprocessing opinionated YouTube videos using Kaldi to segment audio into sentence-aligned subclips, removing silences for clarity. Audio features (e.g., MFCC, chroma, spectral centroid) were extracted with Librosa, while MediaPipe provided frame-level 3D facial landmark coordinates. Sentiment labels from transcribed text were classified using phi 3 mini model. Synchronization between modalities ensured alignment of audio, video, and text features. A Tensor Fusion Network (TFN) integrated these modalities with Conv1D pipelines and custom fusion layers. A late-fusion step incorporated text sentiment predictions. The evaluation employed categorical cross entropy loss, the Adam optimizer, and metrics such as accuracy, precision, and recall for performance tracking. This robust setup emphasized multimodal alignment to capture nuanced sentiment expressions effectively.

V. Results and Discussion

5.1 Experimenting on MOSI and MOSEI Dataset

The first group of experiments sought to test the performance of SLMFN regarding sentiment analysis from opinionated videos from youtube for binary classification. For the Sentiment analysis, we implemented two classes (positive, negative), five classes (Highly positive, positive, neutral, negative, highly negative), classification. Results are there as shown in Table 2 and Table 3 . The datasets we used in the experiments were drawn from MOSEI and MOSI. For the comparison, we also presented the results of several well-known fusion networks [27][28][29][30] on the same dataset as in (see Table 2 AND Table 3). The results were impressive for two classes and for all classification labels (all measurements ranged between 85% and 99%). These results show that combining capabilities of foundation language models and fine tuning them for specific affective computing tasks could improve the classification performance. For example, Unimodal accuracy for MOSEI dataset =

98.05%, while for MOSI = 97.55% . This implies that MOSI slightly degraded the performance of MOSEI. Using the SLMFN is acceptable since recall in binary classification and multi class classification is the same as the accuracy of the classification labels, and thus, higher recall means the method is better able to detect the defined classes i.e. positive, negative and neutral. Also, the fact that improvements occurred in multiclass classification means that such improvements are not coincidental.

5.2 Quantitative Results

The performance of the small language model fusion network was evaluated against certain baseline mode as given in table 2 for multimodal sentiment analysis for CMU-MOSI dataset and in table 3 for CMU-MOSEI dataset. To assess the model performance binary accuracy, f1-score, recall, Mean Absolute Error (MAE), Corr used. MAE is calculated using following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (1)$$

In equ.1 n is the total number of predictions, x_i is the predicted value and y_i is the true value. Corr is calculated using following formula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

In equ.2 r is the correlation, xi is the predicted value, yi is the true value, $\bar{x}$ is mean of predicted value and $\bar{y}$ is mean of true value.

Fusion Method	Model Name	NLP Model Used	2-Class Accuracy	5-Class Accuracy	F1 Score	MAE	Corr
Utterance Based Fusion	**MISA [27]**	BERT	81.78	48.98	81.73	0.761	0.782
Hierarchical Fusion	**MMIM [28]**	BERT	81.49	50.36	81.48	0.743	0.778
Self-Attention Based Fusion	**Self_MM [29]**	BERT	83.23	52.92	83.16	0.715	0.792
Transformer Based Fusion	**MuLT [30]**	Glove	79.81	42.2	79.76	0.923	0.676
SLM Based Fusion	**SLMFN**	Small Language Model	88.3	54.8	88	0.615	0.792

Table 2 Model Performance Comparison on CMU-MOSI Dataset

Fusion Method	Model Name	NLP Model Used	2-Class Accuracy	5-Class Accuracy	F1 Score	MAE	Corr
Utterance Based Fusion	**MISA [27]**	BERT	79.54	53.74	80.28	0.545	0.765
Hierarchical Fusion	**MMIM [28]**	BERT	81.52	53.28	82.42	0.526	0.772
Self-Attention Based Fusion	**Self_MM [29]**	BERT	83.6	55.68	83.56	0.532	0.766
Transformer Based Fusion	**MuLT [30]**	Glove	80.2	53.89	80.76	0.563	0.733

SLM Based Fusion	**SLMFN**	Small Language Model	90.05	61.00	88.00	0.470	0.812

Table 3 Model Performance Comparison on CMU-MOSEI Dataset

5.3 Ablation Studies

To understand the contribution of key components in the small language model fusion network, ablation studies were conducted. The effect of small language model fusion analysed by removing the fine-tuned decoder layers and comparing the performance. Without fine-tuned decoder layers and, MSE increased from 0.470 to 0.0519, and recall dropped from 0.81 to 0.78, highlighting the importance of this component in ensuring smooth sentiment classification. The integration of explainability tools was also evaluated. While it does not directly impact sentiment analysis performance, the interpretability provided by attention heatmaps enhances the usability of the model in multimodal sentiment analysis for research and real-world applications.

VI. Conclusion and Future Work

This work explored the capabilities of multimodal affective computing by integrating the small language model with audio-visual processing techniques for sentiment analysis. Our contributions include methods for aligning and synchronizing characteristics across many modalities, resulting in seamless integration of text, audio, and visual data. We show how small language model can function as a powerful language subnetwork, successfully addressing affective computing tasks such as sentiment analysis and can be also performed on another affective computing task of emotion recognition. In addition, we investigate the utility of small language model's for performing language specific task, demonstrating how they improve both interpretability and classification accuracy. We also offer unique strategies for combining outputs from small language model and an audio-visual subnetwork, experimenting with late fusion method. For ensuring transparency and trustworthiness in affective computing tasks, explainable component in the network will be added in future which enhances interpretability of final classification decision.

References

[1] Soujanya Poria, Erik Cambria, Rajiv Bajpai, Amir Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion, Volume 37, 2017, Pages 98-125, ISSN 1566-2535. https://doi.org/10.1016/j.inffus.2017.02.003.

[2] Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (2017). Affective Computing and Sentiment Analysis. In: Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (eds) A Practical Guide to Sentiment Analysis. Socio-Affective Computing, vol 5. Springer, Cham. https://doi.org/10.1007/978-3-319-55394-8_1.

[3] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE transactions on pattern analysis and machine intelligence 31 (1) (2009) 39–58. https://doi.org/10.1109/TPAMI.2008.52

[4] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

[5] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[6] Liu, Z., Lin, W., Shi, Y., & Zhao, J. (2021, August). A robustly optimized BERT pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics* (pp. 471-484). Cham: Springer International Publishing.

[7] Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems*, *38*(2), 15-23. https://doi.org/10.1109/MIS.2023.3254179

[8] Li, J., Li, H., Pan, Z., & Pan, G. (2023). Prompt ChatGPT In MNER: Improved multimodal named entity recognition method based on auxiliary refining knowledge from ChatGPT. *arXiv preprint arXiv:2305.12212*.

[9] Liu, Z., Yang, K., Xie, Q., Zhang, T., & Ananiadou, S. (2024, August). Emollms: A series of

emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5487-5496).

[10] Zhang, T., Teng, S., Jia, H., & D'Alfonso, S. (2024, October). Leveraging llms to predict affective states via smartphone sensor features. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 709-716).

[11] Broekens, J., Hilpert, B., Verberne, S., Baraka, K., Gebhard, P., & Plaat, A. (2023, September). Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1-8). IEEE.

[12] K. Cortiñas-Lorenzo and G. Lacey, "Toward Explainable Affective Computing: A Review," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 13101-13121, Oct. 2024, https://doi.org/10.1109/TNNLS.2023.3270027

[13] M. M. Amin, R. Mao, E. Cambria and B. W. Schuller, "A Wide Evaluation of ChatGPT on Affective Computing Tasks," in *IEEE Transactions on Affective Computing*, vol. 15, no. 4, pp. 2204-2212, Oct.-Dec. 2024, https://doi.ieeecomputersociety.org/10.1109/TAFFC.2024.3419593

[14] J. Wei et al., "Emergent Abilities of Large Language Models,"https://doi.org/10.48550/arXiv.2206.07682

[15] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[16] Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, *1*, 43-52. https://doi.org/10.1007/s13042-010-0001-0

[17] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (Vol. 1, p. 2).

[18] Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., ... & Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

[19] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

[20] Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018, July). 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)* (pp. 2236-2246. https://doi.org/10.18653/v1/P18-1208.

[21] Gandhi, A., Adhvaryu, K., & Khanduja, V. (2021, December). Multimodal sentiment analysis: review, application domains and future directions. In *2021 IEEE Pune section international conference (PuneCon)* (pp. 1-5).

[22] Ravanelli, M., Parcollet, T., & Bengio, Y. (2019, May). The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6465-6469). IEEE.

[23] Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[24] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In *SciPy* (pp. 18-24).

[25] Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

[26] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

[27] Hazarika, D., Zimmermann, R., & Poria, S. (2020, October). Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia (pp. 1122-1131).

[28] Han, W., Chen, H., & Poria, S. (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

[29] Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021, May). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 10790-10797).

[30] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July).

Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting* (Vol. 2019, p. 6558).
[31] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).