Unified Large-Scale Multimodal Sentiment Analysis Model with Enhanced Generalization

Arata M. Katayama
Earlham College
Richmond, Indiana, USA
amkatay21@earlham.edu

Abstract

Sentiment analysis is a domain of natural language processing (NLP) focusing on interpreting emotions, attitudes, and sentiments primarily from written text. However, recent research has extended this analysis to include audio and visual data, creating a burgeoning field known as Multimodal Sentiment Analysis (MSA). MSA is an evolving discipline dedicated to comprehending emotional expressions and sentiments in not only text but also acoustic and visual inputs. Many MSA models have been built in recent years with different frameworks [6], from statistical non-machine learning techniques to complex neural networks. However, not many focus on generalization performance, which is the ability to predict unprecedented data. This project focuses on using a pre-trained Long Short-Term Memory (LSTM) based network together with Tensor Fusion Network (TFN) and Select-Additive Learning (SAL) algorithm to improve the model's generalizability.

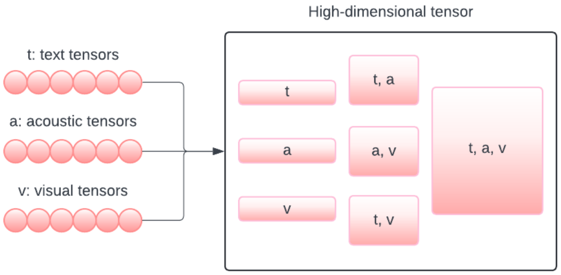

Figure 1: Outer product during tensor fusion

1 Introduction

Sentiment analysis, a crucial aspect of NLP within the domain of artificial intelligence, has traditionally focused on interpreting emotions and attitudes conveyed through written text. However, the evolution of this field has led to the emergence of Multimodal Sentiment Analysis (MSA), extending the analysis to include audio and visual data. The significance of MSA lies in a holistic assessment of emotions and sentiments through multiple modalities, reflecting the multimodal nature of our interactions in the world [6].

While unimodal sentiment analysis is important, particularly for applications like automating customer review analysis [5], textual information alone often falls short of fully capturing the conveyed sentiments. By integrating text, audio, and video, MSA enables the development of more detailed and accurate sentiment analysis models. This advancement extends the applicability of sentiment analysis across diverse domains, from marketing to mental health support [10], where interpreting the full spectrum of human sentiment is integral.

Recently researchers have utilized machine learning (ML) models, such as neural networks like Convolutional Neural Networks (CNN), and Large Language Models (LLM), including GPT [4] and BERT [3]. The integration of these models into MSA frameworks introduces trainable feature extractors for diverse modalities, significantly enhancing the accuracy of MSA models. Additionally, efforts in the MSA field have focused on creating a general model that can effectively analyze sentiments across different video themes and sources. Improving the generalizability of a model enhances its ability to comprehend and analyze sentiments from a wide variety of sources, making it a versatile tool in practical applications.

However, improving the generalizability of MSA models can be hard for a few reasons. One being the limited sizes and ranges of datasets. With limited availability in the ready-to-use multimodal dataset, the range of data a model can train on is automatically limited, making it hard to build a general model. Furthermore, issues with confounding factors are substantial especially in small datasets. Confounding factors are false information which consequently misleads the training process of ML models. For example, Wang et al. illustrate how something as incidental as speakers wearing glasses could become a confounding factor [9]. If all speakers who wear glasses in a dataset express negative sentiments, the model might incorrectly learn to associate glasses with negative sentiments. The smaller the dataset, the more significant the influence and bias of such factors.

This paper specifically focuses on enhancing the generalization performance of MSA models by tackling the presented issues within the MSA field. More specifically, we delve into the various modalities integral to MSA, including text, audio, and visual cues. Our exploration extends to feature extraction methods unique to each modality and the subsequent fusion methods of these features for sentiment predictions. We then discuss the selection of models to enhance generalization capabilities, ultimately proposing a new general MSA model named Select Additive Learning - Tensor Fusion Long Short-Term Memory (SAL-TFLSTM) with the goal of advancing the field of MSA.

2 Related Works

Research in the field of MSA has been growing rapidly with better models and more comprehensive dataset for model training. This paper focuses on two state-of-the-art MSA models which focuses on the aspect of generalizability.

Table 1: Comparison of General MSA models

Model	Fusion Method	Dataset
SAL-CNN	Early / Late Fusion	MOSI, MOUD
Tensor Fusion Network	Early Fusion	MOSI, MOSEI
TransModality	End-2-End	MOSI, IEMOCAP

2.1 Tensor Fusion Network

Multimodal fusion is the process of integrating information from multiple modalities to generate a predictive sentiment score. This fusion can occur at different stages of the modeling process: either at the input phase or at the classification phase. Early Fusion, occurring at the input phase, typically involves a straightforward concatenation of tensors from all modalities. In contrast, Late Fusion takes place at the classification phase, where separate models are used for each modality, and their outputs are combined subsequently.

A notable approach in multimodal fusion is the Tensor Fusion Network (TFN), developed by Zadeh et al. in 2017. [11] This is an Early Fusion method that effectively models both intermodality and intramodality dynamics using Kronecker product, offering a more comprehensive understanding of multimodal interactions. The TFN consists of three substructures: 1) *Modality Embedding Subnetwork*, 2) *Tensor Fusion Layer*, and 3) *Sentiment Inference Subnetwork*. In the original work, the spoken language embedding subnetwork consists of a long short-term memory (LSTM) network with forget gates to learn semantics in a sequential fashion, while the acoustic and visual embedding networks consist of 3 layers of 32 ReLU units. The embedding subnetwork models the intramodality dynamics from the extracted features of each modality. The tensor fusion layer, demonstrated in Figure 1, takes the resulting embeddings of all modalities and performs the Kronecker product (outer product of matrices) to build a high-dimensional tensor ready to be read and analyzed by a machine learning model. In their work using the CMU-MOSI dataset, the TFN achieved the highest accuracy compared to other contemporary state-of-the-art models, demonstrating its effectiveness in handling complex multimodal data.

2.2 Select-Additive Learning

Confounding factors are external variables or influences that can mislead a machine learning model during its training phase. They inadvertently introduce bias, making it challenging for the model to accurately learn the true relationships within the data, as it might attribute effects to the wrong causes. This can significantly compromise the validity of the model's predictions, especially when using a relatively small dataset.

The Select-Additive Learning (SAL) algorithm, introduced by Wang et al. in 2017 [9], addresses these issues of confounding factors. The SAL algorithm is specifically designed to improve the robustness of discriminative neural networks by mitigating the effects of these confounding factors through a two-phase process: the selection phase and the addition phase.

The primary goal of the selection phase is to identify and isolate the dimensions in the original feature representations that are influenced by confounding factors, such as identity traits (e.g., race, ethnicity, or physical attributes like wearing glasses). This is critical in MSA, where such identity features should not affect the sentiment outcome. During this phase, the algorithm employs a specialized loss function that includes a scalar to control the influence of a sparsity regularizer [9]. This helps in effectively pinpointing the confounding dimensions by emphasizing feature selection that contributes genuinely to sentiment analysis while down-tuning irrelevant features.

After isolating the confounding dimensions, the addition phase introduces Gaussian noise, or white noise, to the data. This step simulates random variability or errors in the data, which aids in training the neural network to focus on relevant features by disregarding those confounded by external factors. The addition of noise helps to ensure that the model is trained on a dataset that is essentially free of confounding influences, thereby enhancing the robustness and generalizability of the model.

In the work of Wang et al., the SAL algorithm was optimized for fine-tuning a 7-layer CNN-based network pre-trained on the CMU-MOSI dataset [12] to increase its generalizability.

3 SAL-TFLSTM

The primary goal of this study is to enhance the generalizability of MSA models beyond existing frameworks such as TFN and SAL by merging their methodologies into a comprehensive model called Select-Additive Learning Tensor Fusion LSTM or, simply, SAL-TFLSTM. In this unified model, data modalities are integrated at the input level using tensor fusion, which effectively merges textual, visual, and acoustic data into a cohesive tensor. This tensor is then processed through a bi-directional LSTM, capable of handling variable sequence lengths. Finally, SAL is applied to further fine-tune the parameters of the pre-trained LSTM-based MSA model. This refinement focuses on identifying and adjusting for confounding factors, which would, in turn, improve the model's ability to intricately learn from complex multimodal data and deduce an accurate sentiment prediction.

3.1 Feature Extraction

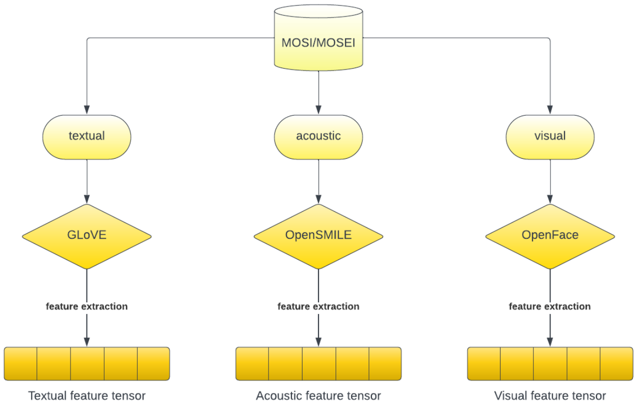

Figure 2: Feature extraction method used in SAL-TFLSTM

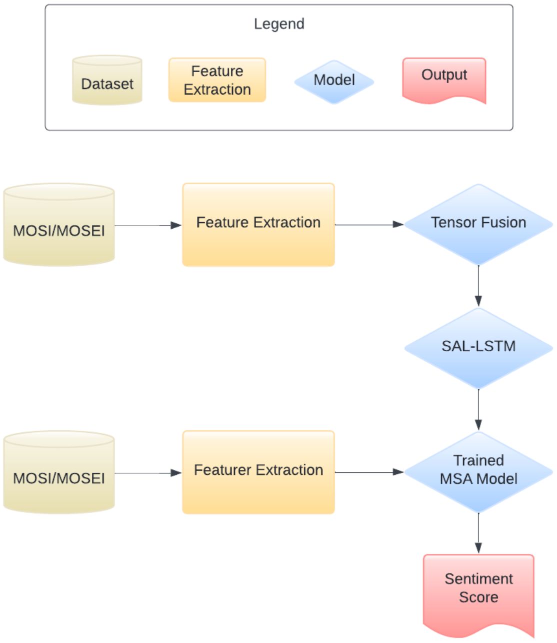

Figure 3: Data architecture diagram of SAL-TFLSTM

3.1.1 *Textual*

For textual feature extraction, we use GLoVE [7]. GLoVE is an unsupervised learning algorithm for generating vector representations of words. SAL-TFLSTM extracts textual features using the GLoVE 300-dimensional word vector, where hence the name, each word vector is represented in a 300-dimensional matrix. These vectors are created by analyzing word co-occurrences in a large corpus. The algorithm constructs a co-occurrence matrix that counts how frequently words appear together within a certain context window in the corpus. This feature extractor precisely captures not just the frequency but also the deeper semantic relationships between words.

3.1.2 *Acoustic*

Collaborative Voice Analysis Repository, or COVAREP, is an open-source voice analysis tool contributed by Degottex et al [2]. It was chosen for acoustic feature extraction because it provides a wide range of algorithms optimized for voice and speech analysis. COVAREP offers tools for fundamental frequency extraction, voicing decision metrics, and detailed glottal source modeling, and many more. These features are essential for high-precision acoustic analysis, which can significantly enhance the capability of MSA models to capture acoustic details. Moreover, the open-source software is built on a collaborative effort ensuring that it continually incorporates updates with the latest advancements in the field.

3.1.3 *Visual*

For visual feature extraction, we use OpenFace [1], an open-source library providing tools for facial expression analysis. This software is particularly suitable for MSA tasks, as it can accurately detect and analyze facial expressions that indicate emotional states, which can be integrated with other data modalities. OpenFace's high accuracy, real-time processing capabilities, and ease of integration with other data streams make it a robust choice for applications that require comprehensive sentiment analysis across different communication channels.

3.2 SAL for LSTM

The SAL algorithm was originally developed for optimizing the learning of CNN based MSA models. However, the issue with CNN based architecture is that it cannot learn the intra-modality dynamics. Since multimodal datasets are extracted from online videos, all modalities are sequential datapoints of a video segment from certain time frame. For this reason, LSTM based architecture seemed to be a more feasible choice. Consequently, SAL needed to be adapted for use with LSTMs. Despite the change in architecture, the core principles of SAL remain unchanged: the prediction of confounding influences and their mitigation. For LSTMs, this adaptation shifts focus from spatial to temporal features within the LSTM's hidden states. The goal is to develop a mechanism that identifies and quantifies the extent to which these hidden states are affected by undesirable biases, such as irrelevant stylistic elements in text, which do not contribute to the task of sentiment analysis.

4 Experiments

The experiments are designed to assess the generalizability of the SAL-TFLSTM compared to state-of-the-art approaches, utilizing a cross-validation method. This involves using separate datasets for training and testing. By evaluating the model's prediction accuracy on data it has not been trained on, we can effectively determine its generalizability to unseen data.

4.1 Datasets

The cross-validation is done with the CMU-MOSI and CMU-MOSEI datasets primarily because of the consistent output metrics in both datasets. Both are annotated with sentiment and emotion intensities on a scale of [-3,3], offering a consistent basis for comparing and validating model performance across distinctive multimodal datasets. The consistency in the output format is important when testing the generalizability of a model, as it ensures that the developed model can be rigorously evaluated under similar standards, ensuring more fair experiment results.

The process of cross-validation involves splitting each dataset into training and testing sets, then training models on one set and validating them on another to assess their predictive accuracy and generalizability. By employing CMU-MOSI and CMU-MOSEI, the models can be exposed to a wider array of video blog data and differing expressions of sentiment and emotion, enhancing their robustness and applicability. This method not only helps in fine-tuning the models to reduce overfitting but also ensures that the learned patterns are not specific to one particular dataset. Such a strategy is particularly effective in research areas like sentiment analysis and emotion recognition, where the ability to accurately interpret and predict across different contexts and modalities is crucial.

4.1.1 *CMU-MOSI*

The CMU-MOSI dataset, as referenced in Zadeh et al. (2016), focuses on sentiment and subjectivity analysis across online opinion videos from YouTube vlogs. This multimodal dataset comprises 3,702 video segments, of which 2,199 are opinion segments specifically annotated for subjectivity and sentiment analysis, and 1,503 are objective segments considered to have neutral sentiment. It features 93 videos with 89 diverse speakers, ensuring a demographically balanced dataset. Each video is manually transcribed and meticulously annotated for both subjectivity and sentiment intensity with a range of [-3,3]. Multimodal data alignment must be ensured as this process allows for an in-depth examination of how different modalities influence the perception and analysis of sentiment and subjectivity in multimedia content.

4.1.2 *CMU-MOSEI*

A larger and more comprehensive dataset is the CMU-MOSEI dataset [13]. Currently the largest multimodal dataset based in English, it offers an extensive resource for the analysis of sentiment and emotion in online opinion videos. It features annotations for 23,453 video segments from 1,000 distinct speakers across 250 topics, providing a rich and diverse collection of multimodal language data. Each segment is manually transcribed and aligned with detailed audio to phoneme levels, encompassing language, visual expressions, and acoustic modalities.

4.2 Evaluation Metric

Since the sentiment labels are floating points within the range [-3,3] for both datasets, regression analysis was the most appropriate. For this reason, the accuracy of the model was measured using Mean Squared Error (MSE) and Mean Average Error (MAE).

4.3 Procedures

The three models, Early Fusion LSTM (EFLSTM), SAL-EFLSTM, and SAL-TFLSTM, are evaluated and compared to assess the impact of the SAL algorithm and tensor fusion method on generalizability. Their performance is analyzed across four different experiments. The initial two experiments focus on measuring the relative accuracy of the MSA models, while the subsequent two aim to determine their generalized accuracy.

- Exp. 1: Trained and tested on CMU-MOSI
- Exp. 2: Trained and tested on CMU-MOSEI
- Exp. 3: Trained on CMU-MOSI, tested on CMU-MOSEI
- Exp. 4: Trained on CMU-MOSEI, tested on CMU-MOSI

5 Experimental Results

5.1 Within Data

Table 2 shows the results of experiments 1 and 2, where all models were trained and tested on different portions of the same dataset. The effectiveness of the SAL algorithm is evident from the MSE values observed when comparing the EFLSTM and SAL-EFLSTM, both trained and tested on the CMU-MOSI dataset. With an output range of [-3, 3], an MSE value of 3.364 for the EFLSTM indicates poor prediction accuracy. However, the introduction of the SAL method reduced the MSE value by nearly a factor of three, underscoring the SAL algorithm's significance, particularly when applied to smaller datasets. The influence of the SAL algorithm was also noticeable on the models trained and tested on the CMU-MOSEI dataset, though the improvements were less apparent than those observed with the earlier model.

Table 2: Within Datasets

Model	Modalities	Metrics	MOSI	MOSEI
EFLSTM	t+a+v	TSP	1.544	1.472
		MSE	3.364	2.913
		MAE	1.544	1.472
SAL-EFLSTM	t+a+v	TSP	1.820	1.820
		MSE	1.249	1.249
		MAE	0.838	0.838
SAL-TFLSTM	t+a+v	-	-	-

5.2 Cross Validation

Table 3 shows the results of experiments 3 and 4, where all models were trained and tested on different datasets.

Table 3: Cross-Validation

Model	Modalities	Metrics	MOSI	MOSEI
EFLSTM	t+a+v	-	-	-
SAL-EFLSTM	t+a+v	-	-	-
SAL-TFLSTM	t+a+v	-	-	-

It was not feasible to conduct cross-validation of the models due to the complexities involved in adjusting model parameter sizes. Consequently, assessing the generalizability of these models is a challenge that will need to be addressed in future research.

6 Future Works

6.1 Integration of Tensor Fusion Network

SAL was originally developed for CNNs, but this study successfully adapted its use for an LSTM-based MSA model. However, the complete integration of the Tensor Fusion Network (TFN) remains unsatisfied. A significant challenge is that TFN generates memory-intensive, high-dimensional tensors, which are difficult to manage. A key area for future research would be to fully implement the tensor fusion model, allowing the SAL-LSTM to utilize the output tensor from tensor fusion as its input instead of relying on simple concatenation.

6.2 Cross-Validation

There are several issues with cross-validation, one being every dataset has a different output format. For instance, in the CMU-MOSI dataset, sentiment outputs range from [-3,3], whereas MELD categorizes outputs into seven distinct emotions [8]. To ensure accurate cross-validation, these output formats need to be standardized or adapted accordingly. Additionally, the input dimensionality of corresponding modalities differs across datasets. Although MOSI and MOSEI are among the most compatible MSA datasets, they still present challenges, such as incompatible numbers of input dimensions for visual features extracted using OpenFace. This discrepancy results in incompatible parameter sizes between models trained on one dataset and tested on another, further complicating effective cross-validation.

7 Acknowledgement

Appreciation extends to Dr. David Barbella for his guidance in planning this research, Dr. Charlie Peck for the support during the execution of the research, Porter Libby for setting up the system environment, and all peers for supporting me throughout this research.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–10.

[2] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30 (2020), 681–694.

[5] Conor Gallagher, Eoghan Furey, and Kevin Curran. 2019. The application of sentiment analysis and text analytics to customer experience reviews to understand what customers are really saying. *International Journal of Data Warehousing and Mining (IJDWM)* 15, 4 (2019), 21–47.

[6] Songning Lai, Haoxuan Xu, Xifeng Hu, Zhaoxia Ren, and Zhi Liu. 2023. Multimodal Sentiment Analysis: A Survey. *arXiv preprint arXiv:2305.07611* (2023).

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[8] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

[9] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 949–954.

[10] Amir Hossein Yazdavar, Mohammad Saeid Mahdavinejad, Goonmeet Bajaj, William Romine, Amit Sheth, Amir Hassan Monadjemi, Krishnaprasad Thirunarayan, John M Meddar, Annie Myers, Jyotishman Pathak, et al. 2020. Multimodal mental health analysis in social media. *Plos one* 15, 4 (2020), e0226248.

[11] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[12] 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* (2016).

[13] 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)*. 2236–2246.