Multimodal Sentiment Analysis

Lucía Prado Fernández-Vega **Andrés Martínez Fuentes** **Javier Ríos Montes**

Pablo García Molina

Abstract

Multi-modal Sentiment Analysis (MSA) has become increasingly important in natural language processing, particularly in understanding user sentiment across various media forms. With recent advancements in deep learning, MSA has been further developed both application-wise and in research, making it a highly explored and relevant field of study. In this paper, we study the challenges and advancements within MSA. Drawing from an extensive research of multi-modal sentiment analysis, our paper introduces two studied model architectures. The first model integrates text and audio data using dual neural networks, aiming to improve the understanding of speech data beyond traditional audio-based classifiers. The second model combines image and text information extracted from the SentiCap dataset using a Transformer Encoder model. We assess the effectiveness of our approaches using the IEMOCAP and SentiCap datasets, along with diverse model architectures. Finally, we express our conclusions and propose future research directions to explore the potential applications of sentiment analysis further.

1 Introduction

Sentiment Analysis is a natural language processing task based on computational methods to determine the emotion of a given source of information, such as a phrase said by a person. In today's digital age, understanding sentiment has become more complex due the incorporation of images, videos, and audio in textual communication, introducing hidden emotions within human expression. Sentiment analysis now requires a shift towards more sophisticated approaches to integrate various modalities of expression, resulting in the development of multi-modal sentiment analysis, to take into account both explicit and implicit expressions.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

MSA goes beyond academic research, with practical applications across various domains(5), including market prediction and business analytics. By integrating data from diverse sources such as social media, news analysis, and product reviews, MSA gives businesses a deeper insight into market trends and consumer behavior to optimize marketing strategies and anticipate shifts in consumer preferences. Further application in this context involves offering personalized recommendations adjusted to individual preferences in recommendation systems to improve user experience and increase sales. In healthcare, it may be used as a tool for mental health prediction by providing insights into emotional well-being and treatment effectiveness through analyzing multi-modal patient feedback. (6) In education, MSA promotes student feedback analysis and the study of emotional dynamics to personalize learning by recognizing students' emotional states and adjusting individual academic content accordingly. Furthermore, MSA is useful in box office prediction, particularly in the entertainment industry. By analyzing sentiment from sources such as social media discussions, reviews, and audience reactions, MSA helps studios forecast the success of upcoming releases, allowing stakeholders to allocate resources effectively and maximize box office revenues.

However, achieving effective multi-modal sentiment analysis faces significant challenges. Integrating diverse modalities introduces technical obstacles, such as data fusion, feature extraction, and model design, especially when aiming to detect hidden emotions such as sarcasm, contextual and complex sentiments. Moreover, the field encounters obstacles related to dataset availability: scarcity of large, diverse datasets with high annotation accuracy, along with difficulties in analyzing video data due to noise and low-resolution footage, persist as unresolved issues.

In this paper, we research the difficulties of multi-modal sentiment analysis, focusing on speech and image emotion recognition. Our study aims evaluate the ability of different model architectures to capture emotional features found in different modalities of data, which may not be fully captured by analyzing text only. We build upon existing methods in MSA, such as BERT for text processing, attention mechanisms for audio analysis and Res-Net for image classification and address key challenges such as dataset selection, modality misalignment, model complexity and sentiment discrepancy. In the following sections of this paper, we explain the methodology employed in our study, present our experimental findings, and discuss their implications.

2 Prior Related Work

Sentiment analysis first appeared as a term in the early 2000s, together with other similar concepts such as opinion mining, despite having deep roots on public opinion analysis at the start of the 20th century (8), being in hibernation until the availability of both sufficient amounts of data and models. Example of this is (9), a 2003 study in which online reviews were analysed, using a Naive Bayes Classifier in order to assess their positivity or negativity. Through the early years, the publishing of papers on this matter multiplied fast, as better models (such as LSTMs and Deep Networks) and availability of greater datasets (specially with the advent of social media) allowed insights to be drawn in a far easier way. In the last decade, great advances have been accomplished in the matter, greatly due to the designing of new techniques, such as Attention mechanisms and, lately, Transformers.

Multimodal Sentiment Analysis typically relies on a combination of three modalities: text, audio and visual. Each one contributes different features that allow more accurate predictions to be done. Text, for instance, is the most dominant and common one, though insufficient in instances where cultural factors, use of sarcasm or irony are present. Visual features, on the other hand, help in the identification of the sentiment by characteristics such as the body language and identification of unclear words (polysemy, for example). Last, acoustic features provide additional information by means, for instance, of the tone. This is why MSA is interesting, as it aims at using all the different information present in the modalities in a single model to achieve better results. As the results show, employing a combination of the three modalities in either a bimodal (using two of them) or trimodal (using them all) system guarantees a superior accuracy compared to unimodal system.

Even though it's not the main topic covered in this paper, the fusion of the features present in each modality is a crucial part of the processing (6). MSA fusion techniques are the processes of filtering, extracting and combining all the features received from a variety of sources. Out of the research we've done, three main methods are used: Feature, Decision and Attention-based.

Feature Fusion, also known as early fusion, combines the features present in each modality before feeding it to the classification algorithm. This allows it to take into account, at least partially, the correlation between the different modalities, potentially leading to better results. Despite this, time

synchronisation is necessary for this method, which isn't easy most of the times, and takes a lot of preprocessing in order to work. Added to this, the differences in information between modalities make it difficult to have a universal algorithm to create the join feature vectors. Examples of models using this approach are the THMM (Tri-modal Hidden Markov Model) or RMFN (Recurrent Multistage Fusion Network). These, even though achieving high accuracy and robustness, depend on a large amount of training data to achieve such performance. What's more, their structures are complex, which requires a longer training time.

Decision-level Fusion, on the other hand, is commonly known as late fusion, as the features of each modality are independently processed and classified. Then, results are fused to form a decision vector. This integration is done via different methods, such as averaging, majority voting or learnable models (7). As such, the models using this technique are typically more lightweight and flexible, as when some modality is missing, the decision can be made using the other modalities. Furthermore, given that each modality's features are processed separately, different models can be applied to each one, using the one with the best performance.

Last, attentional feature fusion, proposed in 2021, leverages the concept of attention mechanisms (specifically, cross-modal), dynamically weighing the importance of the information present in each modality (3). This allows the models to be far more flexible than the previous two, prioritising more relevant utterances when needed on one or other modality.

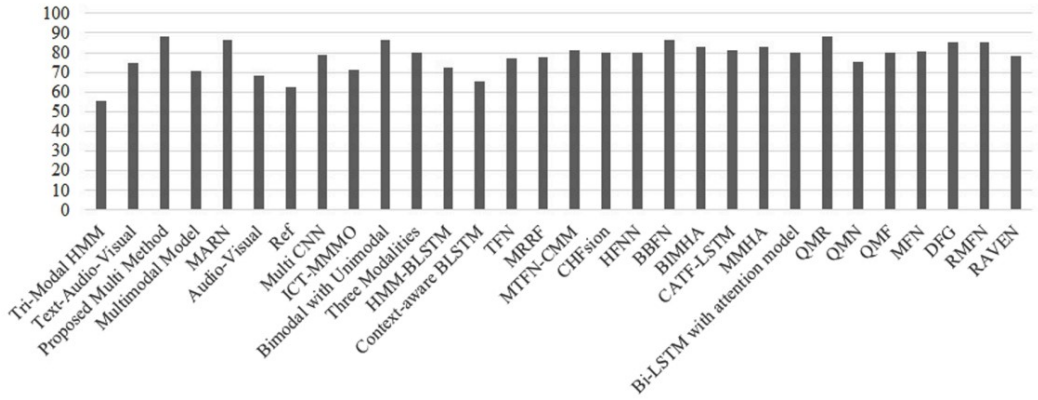


Figure 1: Binary accuracy of different MSA Architectures

In the figure below, we can observe the accuracy different models achieve on a same task. Obtained from (6), it shows how there isn't a given architecture that's been established as supreme with respect to the rest. It can be seen how more novel architectures, like *Bi-LSTM with attention model* from 2021, can perform worse than other, arguably, simpler ones, like *Bimodal with Unimodal* from 2015. This implies that there are many additional factors apart from the complexity or novelty of the architecture.

3 Models

This section describes the architecture and methods of our proposed models for MSA. Our approach focuses on emotion recognition in speech and images, integrating different processing modules for each modality. Our proposal consist of two main architectures: the Audio-Text (AAT) model and the Output Transformer Encoder (OTE) model.

3.1 Audio and Text Model

The Audio-Text (AAT) model consists of three main components to perform accurate sentiment analysis in spoken dialogue: audio processing with Alex-Net, text processing with BERT embeddings and feature fusion through a fully connected layer.

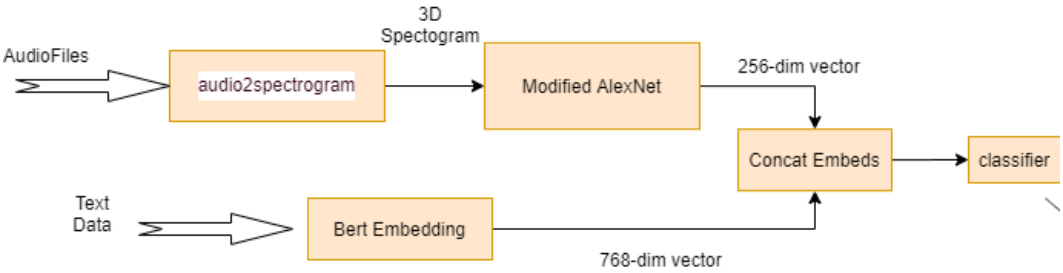


Figure 2: AAT Model Architecture

3.1.1 Alex-Net for Audio

This section details the method employed to convert raw audio data into meaningful representations for sentiment analysis. Our AAT model extracts relevant features from audio inputs by converting to spectrogram representations and passing them through a modified version of the Alex-Net architecture. By applying a logarithmic transformation to the spectrograms, we enhance the contrast and better capture the properties of audio signals. Finally, we introduce an attention mechanism to condense the extracted features into a representative tensor.

3.1.2 Alex-Net

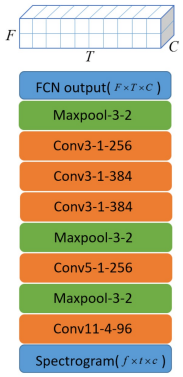


Figure 3: AlexNet Architecture. Convolutional layer parameters are denoted as Conv[kernel]-[stride]-[channels]. The maxpooling layer parameters are denoted as Maxpool-[kernel]-[stride]. The local response normalization layer and ReLU activation function is not shown.

Our audio processing module is based in the Alex-Net Fully Convolutional Network (FCN). Inspired by the seminal AlexNet architecture (12), our FCN architecture is designed specifically to process spectrogram representations of audio data.

The FCN architecture is composed of a sequence of convolutional layers, combined with non-linear activation functions such as ReLU, in order to capture complex features within the spectrogram data. Through successive convolutions, the FCN gradually captures details and hierarchical representations from the audio input, simplifying the audio data into manageable representations with a comprehensive representation of the input, characterized by its spatial dimensions (frequency and time) and channel depth.

3.1.3 Attention Layer

Following the extraction of features by the AlexNet FCN, we introduce an attention mechanism to extract the features from the audio data with a greater influence on classification. This serves as a selective mechanism to focus on key regions of interest within the feature space and suppress noise and irrelevant information by assigning varying degrees of importance to different elements of the feature tensor in a condensed representation of the feature vector with the most informative aspects of the input. We use the following formulas to realize this idea:

$$\mathbf{e}_i = \mathbf{u}^\top \tanh(\mathbf{W}\mathbf{a}_i + \mathbf{b}) \tag{1}$$

$$\alpha_i = \frac{\exp(\lambda e_i)}{\sum_{k=1}^{L} \exp(\lambda e_k)} \tag{2}$$

$$\mathbf{c} = \sum_{i=1}^{L} \alpha_i \mathbf{a}_i \tag{3}$$

Through the integration of the attention layer, our model gains the capability to dynamically adapt its focus based on the context of the audio input. However, due to the limitation of computational resources, our network offers an alternative method for further audio processing: a linear module flattening the spectrograms for direct vectorization.

3.1.4 BERT Embeddings for Text

The text processing module is a critical component to convert raw text into meaningful representations that downstream models can understand and use effectively. In this module, we use the features of pre-trained BERT (Bidirectional Encoder Representations from Transformers) embeddings, a recent development in transfer learning for NLP tasks.

BERT provides an efficient method for pretraining language representations, which can be fine-tuned or used to extract features for downstream tasks. The embeddings generated by the pre-trained BERT model from the transformers library by Hugging Face contain useful contextual information about the input text for the extraction of high-quality language features. (10)

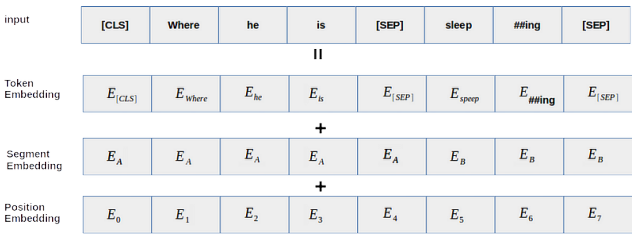


Figure 4: BERT input representation.

Raw text data is tokenized using the BERT tokenizer, which splits the text into individual tokens and converts them into numerical representations for further classification. Unlike traditional word-based tokenization, BERT uses sub-word tokenization, which allows for greater flexibility and effectiveness, specially when handling complex word structures. One of the main sub-word tokenization techniques used in BERT is Byte-Pair Encoding (BPE), which divides text into sub-word units, such as prefixes, suffixes, and root words to effectively represent a wide range of words and word variations. Additionally, BERT tokenization involves the use of special tokens like '[CLS]' (classification), '[SEP]' (separator), and '[MASK]' (masked), for specific purposes in model input. BERT tokenization can

be accessed through the Hugging Face Transformers library, in order to encode text data into input features that the BERT model can understand and process effectively. (1)

Once tokenized, the text is passed through the pre-trained BERT model to obtain embeddings for each token. Unlike traditional word embeddings, which assign a fixed representation to each word regardless of context, BERT embeddings consider surrounding words to capture contextual information dynamically, representing the semantics and relationships within the text more accurately. Finally, the resultant tensor [batch, sequence size, embedding dim] is passed through a fully connected (FC) layer to reduce dimensionality and return a processed representation of the text with dimensions [batch, text embedding size].

3.1.5 AAT Feature Fusion

Finally, the outputs from the different modality processing modules are concatenated and the combined feature representation is then fed into a fully connected (FC) layer with dropout regularization for classification. The FC layer processes the fused features to produce the final logits vector, encapsulating the model's prediction across different sentiment classes.

3.2 Output Transformer Encoder Model

The Output Transformer Encoder (OTE) consists of three main components for emotion recognition in visual data: image processing through a ResNet module, text processing with BERT embeddings and feature fusion through a FC layer, similar to the AAT model.

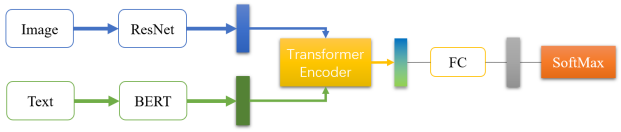


Figure 5: OTE Model Architecture

3.2.1 ResNet-50 for Images

The image processing module employs a Residual Neural Network (ResNet) architecture, specifically ResNet-50, a deep convolutional neural network (CNN) for sequential processing of the input images through a series of residual blocks, each designed to capture increasingly abstract features and facilitate the flow of gradients during training. The ResNet-50 architecture is composed of three key components:

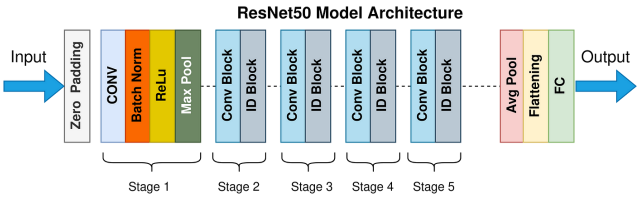


Figure 6: ResNet-50 Architecture

Data Ingestion: Once loaded, the images are resized to a standard size of (64, 64) pixels, maintaining aspect ratio, and converted into PyTorch tensors to ensure uniformity in image dimensions and data type across the dataset. The image tensors are first processed by the data ingestion block, which consists of convolutional and pooling layers, to prepare the images for further processing by extracting low-level features.

Data Processing: The main body of the ResNet-50 architecture is the data processing block, which consists of four modules, each containing several residual blocks. The residual blocks are the building blocks of ResNet and enable the training of very deep networks. These blocks consist of two branches: the convolutional branch and the residual branch. The convolutional branch applies a series of convolutional layers with batch normalization and ReLU activation functions to extract

features from the input. The residual branch performs identity mapping, passing the input forward unchanged. If the input and output channels of the residual block are different, a 1x1 convolutional layer is applied to the residual branch to match the dimensions. The output of both branches is then added together element-wise, and the result is passed through a ReLU activation function. This process allows the network to learn residual mappings, making it easier to optimize deeper networks.

Prediction: The output features are processed through a prediction block, which consists of an adaptive average pooling layer followed by a FC linear layer. The adaptive average pooling layer aggregates spatial information from the feature maps, and the FC linear layer maps the features to the desired number of output classes, producing the final logits tensor.

The ResNet50 class is a wrapper that encapsulates the entire ResNet-50 architecture. It takes as input the number of input channels and the desired output dimension and constructs the ResNet-50 model accordingly.

3.2.2 BERT Embeddings for Text

Similar to the AAT model, the text processing module uses the pre-trained BERT model to obtain embeddings for each token. However, due to the limitation of computational resources, the OTE model includes the Word2Vec-type-based embedding module as an alternative method to BERT embeddings, which is computationally less demanding due to its simpler architecture and smaller parameter count. This simplification may limit the models capacity to capture contextual information.

3.2.3 OTE Feature Fusion

In the Output Transformer Encoder (OTE) model, the fusion of image and text features is performed by a Transformer Encoder layer followed by a classifier module. The Transformer Encoder layer operates on the concatenated vector to integrate features from the different modalities and generate comprehensive representations for subsequent sentiment classification. The resulting output tensor undergoes dimensionality reduction to be passed through a classifier module composed of dropout layers and fully connected (FC) layers, which produces the logits tensor representing the model's predictions across different sentiment classes.

4 Data

4.1 IEMOCAP

The IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset serves as the foundation for the AAT model. This dataset consists of 151 videos containing recorded dialogues, with two speakers per session, annotated for the presence of ten distinct emotions (neutral state, frustration, anger, sadness, happiness, excited, surprise, fear, disgust, or other), along with annotations for valence, arousal, and dominance. For the adquisition of the dataset, we drew upon the original database collected at SAIL lab at USC (2).

Deep preprocessing had to be applied to the dataset, in particular segmenting text into utterances, as the original files (.txt) were structured into dialogs, without specifically dividing it up into the statements and pairing them with their respective audios. Additionally, considering that the model takes tensors and spectrograms, the data had to be correctly formatted, reason why we created our own *IEMOCAP_Dataset* class which, for each utterance, returns the matching tensors.

4.2 SentiCap

The SentiCap dataset contains several thousand images with captions with positive and negative sentiments. These sentimental captions are constructed by the authors by re-writing factual descriptions. In total there are 20000+ sentimental captions. The data consists of images and captions referring to them, the number of images is limited so a few captions are provided for each image resulting in a greater number of data. Specifically, for each image there are 5 positive captions and 5 negative ones. This, as specified in the next section, has some effects on the training of individual modalities.

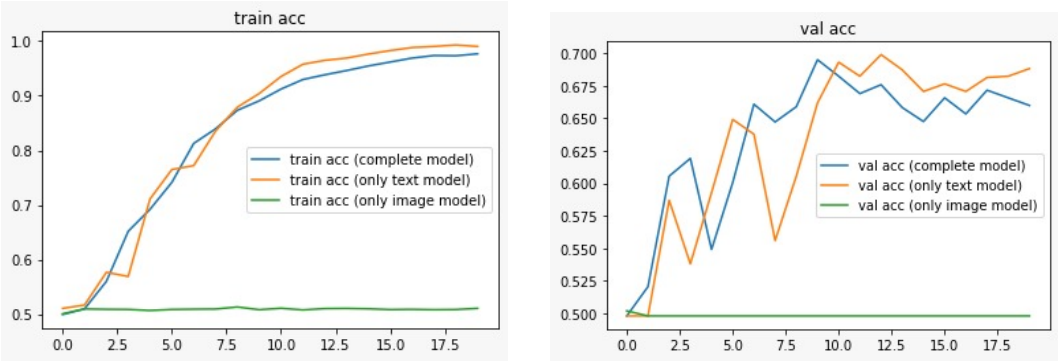
5 Experiments & Results

Our main goal when approaching the experiments was assess the relative performance of each modality and the combination of them. For that purpose, we will first focus on the results we had using individual modalities, and then we will explore the models in their entirety.

5.1 Single modality overview

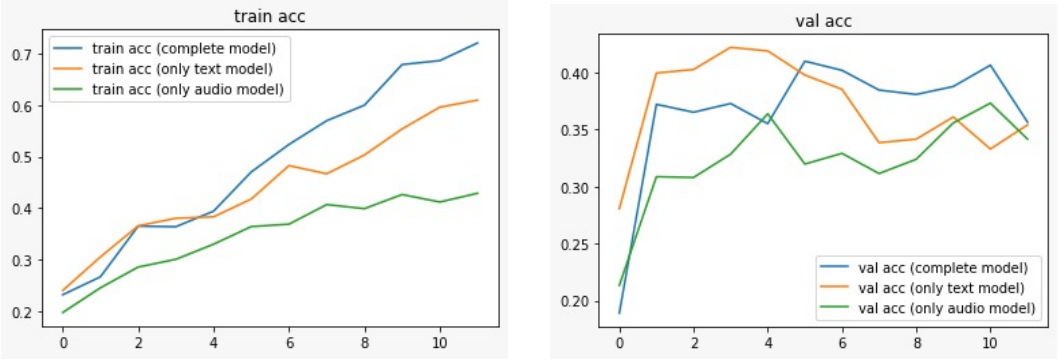
5.1.1 Text-Only

In this case, we use both our models, isolating the text model to assess its relative performance. As it was previously stated in the Section 3.2, the text-processing part of the models depend on Word2Vec embeddings.



(a) Train accuracy (b) Validation accuracy

Figure 7: Train and Validation Accuracies for OTE



(a) Train accuracy (b) Validation accuracy

Figure 8: Train and Validation Accuracies for AAT

As in other sections, we must take into account that OTE performs a binary classification, while AAT must classify under 10 possible classes (therefore, the difference in accuracies). We'll focus on the relative accuracies, inside each model, with the rest of the modalities. In both, text performs relatively similar to its bimodal counterparts though, in AAT, the complete model outperforms it by a bit. To understand what this implies (and the conclusions we might extract of it), we hypothesize that, for sentiment analysis, text is more than enough to, more or less, correctly classify a dialogue, and images don't add much crucial information. Despite of this, in more complex problems (like emotion analysis, the case of AAT), other modalities (audio, in this case) do really matter, as text may not be enough to differ between frustation, sadness or anger, for instance.

This lines up with findings we have collected. For example, as we can see in the table below, while text scores similar results to bimodal or trimodal architecture in binary classification, more complex models (which combine different modalities) tend to outperform text-only as more classes are incolved in the classification. The table, extracted from (11), shows how trimodal architectures are better than other simpler ones, in a magnitude which only increases for multiple-class problems.

Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
$\text{TFN}_{language}$	74.8	75.6	38.5	0.99	0.61
TFN_{visual}	66.8	70.4	30.4	1.13	0.48
$\text{TFN}_{acoustic}$	65.1	67.3	27.5	1.23	0.36
$\text{TFN}_{bimodal}$	75.2	76.0	39.6	0.92	0.65
$\text{TFN}_{trimodal}$	74.5	75.0	38.9	0.93	0.65
$\text{TFN}_{notrimodal}$	75.3	76.2	39.7	0.919	0.66
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
TFN_{early}	75.2	76.2	39.0	0.96	0.63

Figure 9: Results from the ablation study on TFN

5.1.2 Image-only

In this section of our ablation study, we evaluate the contribution of image data to our OTE model's performance in sentiment analysis.

Studying the data in Figure 7b, we can conclude that the inclusion of image data does not improve the model's accuracy; instead, it has no significant impact. During training, the model's accuracy remains consistently flat at 0.5 when using image data alone. Furthermore, the model achieves better performance using only text data compared to a combination of image and text data for prediction. This discrepancy can be attributed to several factors, such as the complexity of understanding visual sentiment tokens, insufficient image quality due to the dimension reduction, or an inadequate model architecture for visual data. Nevertheless, we believe the key issue may be the biased and limited structure of the SentiCap dataset, as it contains repeated images associated with different sentiments. This repetition makes the learning process inconsistent, as the model finds identical visual inputs linked to varying labels.

For a further analysis of the model's performance, we bench-marked our results against those of the following state-of-the-art methodologies in visual sentiment analysis:

3DCNN: A 3D Convolutional Neural Network (CNN) trained using facial data from a speaker (REF 1)

CNN-LSTM: A hybrid model combining CNNs and Long Short-Term Memeory (LSTM) networks, performing convolutions on facial regions at each timestamp and passing the output through an LSTM. (REF 2)

LSTM-FA: A model using information extracted by FACET every six frames as input to an LSTM with a memory dimension of 100 neurons.

Visual Unimodal		
Baseline	Binary Accuracy	5-class
3D-CNN	56.1	24.9
CNN-LSTM	60.7	25.1
LSTM-FA	62.1	26.2
ResNet-50	50.1	-

Table 1: Comparison with state-of-the-art approaches for visual sentiment analysis.

The analysis highlights that using only image data, particularly with the SentiCap dataset, does not improve the model's performance in sentiment analysis. This is likely due to the presence of repeated

images associated with different sentiments, leading to inaccurate predictions. However, more broadly, the ablation study shows that models that rely only on visual data generally perform poorly, as visual data alone is often insufficient for accurately determining sentiment without contextual information provided by text.

5.1.3 Audio-only

In this section of the study, we evaluate the impact of using audio data on our CNN-based model combined with attention mechanisms for sentiment analysis.

Studying Figure 8, we observe that when using audio data only, the model shows a slight improvement, reaching an accuracy of 0.36. Although this shows an improved impact compared to the image-only approach, it remains lower than the accuracy achieved using text data alone. However, when combining audio with text data, the model's performance further improves, proving the audio modality is useful for sentiment analysis. The following factors may explain these observations:

1. Audio provides complementary information not captured by text alone, such as tone and pitch, which are crucial for detecting sentiment keys.

2. The attention mechanism in combination with CNN may be effectively extracting and processing the features from audio data that are indicative of emotions.

3. The combined architecture of CNN and attention mechanisms might be more suited for processing and integrating audio features compared to our architecture for visual features.

Overall, the integration of text and audio allows the model to exploit the strengths of both modalities, resulting in better overall performance. This multimodal approach helps to capture a more precise representation of sentiment by combining linguistic and auditory elements.

For a further analysis, we compared the performance of our AAT model using exclusively audio with the following well-known approaches in auditory sentiment analysis:

HL-RNN (REF 3) uses an LSTM on high-level audio features.

Adieu-Net (REF 4) is an end-to-end approach using PCM features directly.

SER-LSTM (REF 5) uses recurrent neural networks on top of convolution operations on spectrogram of audio.

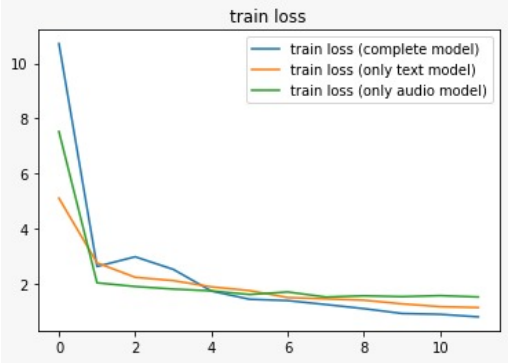
Audio Unimodal		
Baseline	5-class	10-class
HL-RNN	25.9	-
Adieu-Net	25.1	-
SER-LSTM	24.1	-
ResNet-50	-	36.2

Table 2: Comparison with state-of-the-art approaches for acoustic sentiment analysis.

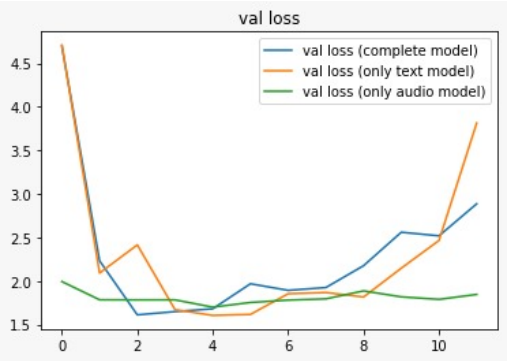
5.2 Full model results

5.2.1 AAT Model

We have already discussed the AAT model in its single modality capabilities, moreover, we can appreciate in the results showing its different accuracies 8 that they both contribute to the performance of the complete model. In any way, the results of the AAT perfectly reflect what could be expected of a MSA model. However, one thing still screeches, the performance of the model in validation barely scratches 40%, whereas the accuracy in training easily reaches 70%.



(a) Train loss graph for AAT model



(b) Validation loss graph for AAT model

Figure 10: Train and Validation loss graphs for AAT model

Just by looking twice at the graphs (10) one can deduce a clear case of overfitting. From the begining, the architecture proposed by the AAT model proved to be insufficient for extracting the relations of the IEMOCAP datset. Even with our best efforts, no combination of hyperparameters showed any improvements from this results. Going further, we were forced to reduce the complexity of the model, once again. For instance, the main issue that was holding down the training for the model was the tremendous size of the audio inputs. For that, we had to reduce the time scale, which also meant to make some audios shorter than they originally were, losing information on the process. Also, we decided to change the AlexNet for a simpler small CNN, since the AlexNet required too much memory, and keep the Attention layer, but this last did not make any difference and could be replaced with a simple Linear layer. This changes strongly changed the data and shrunk the latent spaces of the model.

Hyperparameter	**Value**
`epochs`	12
`lr`	0.01
`batch_size`	64
`dropout`	0.4
`C`	256
`lrn_mode`	full
`lambd`	0.3
`time_dim`	750
`out_text_dim`	256

Table 3: Final overview for the complete hyperparameters that showed better performance

With all in hand, we can state with certainty that the AAT model does not have a proper architecture is not able to generalise and understand audio and text inputs fully. However, it does serve as an initial step forward, since its capable of merging both modalities to obtain a better estimate.

5.2.2 OTE Model

Initially, we weren't able to train the OTE model, using the SentiCap dateset and our available computational resources, to obtain results better than 50% accuracy, which corresponds, basically, to random guessing. Further inspection of the model showed no flaws on the modules of the model, so we had to consider other possibilities. It's important to mention that we had several issues related to insufficient memory capacity before, and we tried adapting the model by further optimizing the way data was loaded at each iteration and different CNN architectures for interpreting image input. In the end, we concluded that the model was not able to capture the behaviour of the data, since it was too complex for our dataset. By simplifying the CNN architecture and lowering the dimensionality of the output spaces from both image and text processing, we were able to make the model learn the patterns.

As a reflection, we discovered that it was not a problem of vanishing gradients, nor the fault of our computers that make the training process tremendously tedious, but rather a matter of critically thinking about the data we had at our disposal and realizing we were blinded by a problem we could have addressed in a much simpler way. Diving deeper into the technical details, we mantained the BERT embedding, although we had a *word2vec* prepared and we also tried experimenting with it, but we changed the deep ResNet-50 with a simple CNN with three convolutions, a pooling operation and some normalizations, which fitted much better our data. Both the image processing and the text processing had an output embedding dimension controlled by two hyperparameters of the model, and the dimension of the parameters of the classifier grows with their multiplication. This last detail turned out to be crucial, by lowering both embedding dimensions we obtained a much lower dimension for the linear classifier, since the matrix for the weights parameters decreased in size, thus made the model simpler and able to train meaningfully. This last change turned out to be the most significant, because only changing the CNN architecture was not sufficient.

Hyperparameter	**Value**
`epochs`	20
`lr`	0.001
`batch_size`	64
`dropout`	0.4
`image_out_dim`	64
`text_out_dim`	64
`classification_hidden_size`	20
`attention_heads`	4
`use_small_cnn`	True
`dim_feed_forward`	2048
`use_word2vec`	False
`text_sequence_max_length`	60

Table 4: Final overview for the complete hyperparamenters that showed better performance

Once these changes were applied, we were able to obtain higher accuracy results of around 70%. Other MSA models that process image and text inputs, with even more complex architectures, obtain similar results. The most used dataset for this two modalities is MVSA. In the study "Multimodal Sentiment Analysis With Image-Text Interaction Network" (13), the authors compare different architectures with theirs, which far exceeds our simple network. Of course, this results can not be taken as a literal comparison, since they were obtained with a different dataset, but they do serve as a guide of how well other models perform.

Method	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
SentiBank&Strength [13]	0.5205	0.5008	0.6562	0.5536
CNN-Multi [57]	0.6120	0.5837	0.6639	0.6419
DNN-LR [58]	0.6142	0.6103	0.6786	0.6633
HSAN [6]	-	0.6690	-	0.6776
MultiSentiNet [35]	0.6984	0.6963	0.6886	0.6811
CoMN [7]	0.7051	0.7001	0.6992	0.6983
MVAN [59]	0.7298	0.7298	0.7236	0.7230
ITIN (Ours)	**0.7519**	**0.7497**	**0.7352**	**0.7349**

Figure 11: Results of different models on the MVSA dataset

Finally, for comparison, we now must comment on the differences between single and multiple modalities. In the figure above [7], we observe similar results are obtained when using only textual data and when using both text and image inputs. This was explained earlier in the discussion of only-image inputs, it is due to the irregularities of the labeled images and its repetitions that the image inputs do not provide useful data for the model. This last reason also explains why the only-text and the full model behave the same, simply the image data does not contribute, nor positively or negatively, to the performance. Is important to highlight that only by looking at the graphs one might

think the only-text implementation behaves better than the full model, however, based on reasons explained just now, we can confidently state that both cases give the same results, since only the textual inputs contributes to the results. Notably, for this dataset, a text-only model would have been more than sufficient.

5.3 Strengths and Weaknesses

In the first place, the main strength of our model is the text processing module as the text is what mainly determines the sentiment of the data because it encapsulates a lot of information. Although our models are simple the performance in both cases is quite good, reaching a 70% in the image-and-text model and 40% in the audio-and-text model.

The main weakness of the model is the sensitivity related to the data, this affects primary in the OTE model as its discussed in the only image subsection, the Senticap Dataset consists in several reviews over different things, the problem is that for each item there are the same amount of positive and negative reviews, that way the model can't learn anything from the image input. This problem on the other hand does not occur in the AAT model as it is discussed in the only audio subsection the audio takes an important role on the complete model structure.

Another weakness is the simplicity of the model, even though the accuracies reached are acceptable, with an architecture more complex and more data the model could learn the relationship between the text input and the audio or image input in order to classify correctly when there is a discrepancy between the two inputs.

6 Analysis & Conclusion

All being said, it is time to clarify our final impressions after the research. First of all, it is clear this field demands great computational power, and deserves time and pamper in order to obtain a feasible solution in a successful and optimal way. Also, one must not forget the proper nature of the problem, since classifying something as abstract and subjective as an emotion is indeed an intricate and difficult task, even for us humans. Both of these conditions position MSA as field still under investigation from a Deep Learning point of view, and one that still does not present a clear solution.

Following the previous research and our own experience, we coincide that at the heart of MSA lies feature fusion. The only way a model can really take advantage of its multi-modal nature is by properly addressing the way it interprets the different inputs. Along this lines, new architectures are being proposed every year, all of them with huge amounts of research and sense to them, each more complex than the last. One clear example is the HCIL architecture (4) (Figure 12), that proposes several modules to tackle and learn different combinations of data and has served as a huge inspiration for our work. However, as for today, there has been no clear winner as to this matter, but one thing is clear, if we are able to find a way to make a neural network learn these relationships, we will be able to solve the most delicate problem in MSA, feature discrepancy.

All our experimental results prove our theoretical research, and show two clear outcomes: the importance of data quality and the need for an adaptable and complex architecture. On the one hand, data has proven to have a huge impact on the results. As such, if the different emotions are not properly reflected on the modalities' data (uniquely identifying the sentiment they are associated with), the results will not be conclusive regardless of how good the model is and there will be little chance for generalization. On the other hand, the choice of the architecture is crucial, still no network has raised as the indisputable key for sentiment analysis, but it surely would have to be able extract the intertwined relations between modalities. Yet again, even though better datasets are being put together over time, MSA stands unbeaten waiting for a new architecture able to perfectly learn and recognise human emotions (if such task is even possible).

Finally, we would like to express our thoughts as to where the efforts should be pointed, and propose a new path for investigation. Once understood the importance of feature fusion, we believe a proper approach to this problem would consist on mapping the inputs between one another in the time dimension. This way, the model will be able to link the word with its sound clues and with the visual impressions it implies as well. It can be seen as an attempt to wrap simultaneously both the issues mentioned in the previous paragraph, since it could be carried out either by improving the input data or by expanding the data processing module of the model. For us, this method would further help

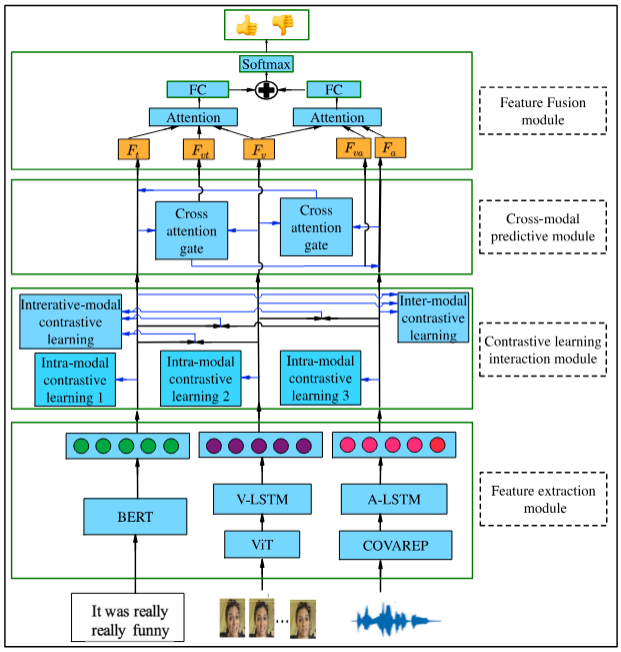


Figure 12: HCIL Model Architecture

to deal with feature discrepancy by learning the moments where different types of inputs appear to disagree and understanding them not as an error but rather as a different emotion.

References

[1] ASGHARPOOR GOLROUDBARI, A. Training your own bert model from scratch.

[2] BUSSO, C., BULUT, M., LEE, C.-C., KAZEMZADEH, A., MOWER, E., KIM, S., CHANG, J. N., LEE, S., AND NARAYANAN, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation 42* (2008), 335–359.

[3] DAI, Y., GIESEKE, F., OEHMCKE, S., WU, Y., AND BARNARD, K. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (2021), pp. 3560–3569.

[4] FU, Y., ZHANG, Z., YANG, R., AND YAO, C. Hybrid cross-modal interaction learning for multimodal sentiment analysis. *Neurocomputing 571* (2024), 127201.

[5] GANDHI, A., ADHVARYU, K., AND KHANDUJA, V. Multimodal sentiment analysis: Review, application domains and future directions. In *2021 IEEE Pune Section International Conference (PuneCon)* (2021), pp. 1–5.

[6] GANDHI, A., ADHVARYU, K., PORIA, S., CAMBRIA, E., AND HUSSAIN, A. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion 91* (2023), 424–444.

[7] KAUR, R., AND KAUTISH, S. Multimodal sentiment analysis: A survey and comparison. *Research anthology on implementing sentiment analysis across multiple disciplines* (2022), 1846–1870.

[8] MÄNTYLÄ, M. V., GRAZIOTIN, D., AND KUUTILA, M. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review 27* (2018), 16–32.

[9] NASUKAWA, T., AND YI, J. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture* (2003), pp. 70–77.

[10] PRAKASH, P. An explanatory guide to bert tokenizer. *Data Science Blogathon* (Sep 2021). Published as part of the Data Science Blogathon.

[11] ZADEH, A., CHEN, M., PORIA, S., CAMBRIA, E., AND MORENCY, L.-P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[12] ZHANG, Y., DU, J., WANG, Z., ZHANG, J., AND TU, Y. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (2018), IEEE, pp. 1771–1775.

[13] ZHU, T., LI, L., YANG, J., ZHAO, S., LIU, H., AND QIAN, J. Multimodal sentiment analysis with image-text interaction network. *IEEE transactions on multimedia* (2022).