Towards Interpretable Multimodal Models for Emotion Recognition

Author:
Kathleen Koosje de Boer
Utrecht University
6439608
k.k.deboer@students.uu.nl

Supervisor:
Dr. Marijn Schraagen

Second Supervisor:
Dr. Albert Gatt

Supervisor Sound & Vision:
Dr. Willemien Sanders

Second Supervisor Sound & Vision:
Rana Klein MSc

Master Artificial Intelligence

Department of Information and Computing Sciences
Utrecht University
Utrecht, The Netherlands
July 3, 2024

Acknowledgements

I would like to thank the people who helped me in various ways throughout the creation of this thesis. First of all, I am grateful to my supervisor, Dr. Marijn Schraagen, for his support, constructive criticism, and insightful comments. I also wish to thank my second supervisor, Dr. Albert Gatt, for his expertise and suggestions, which were helpful in narrowing down the research focus. Special thanks to the people at Sound & Vision, Willemien Sanders, Sara Veldhoen, Rana Klein, Mari Willems, and Jaap Blom for their great collaboration and support. I would also like to thank Dragos Balan, Marta Espanos Lopez, and Maddalena Ghiotto for their support and the enjoyable moments that we shared. Finally, I would like to thank my parents, my sister Hella and Jivan, for their support and love. Your belief in me has been a constant source of motivation.

Sincerely, Thank you.

Abstract

The contents of this thesis focus on the development and evaluation of an interpretable multimodal model for emotion recognition in collaboration with the Dutch Institute of Sound & Vision. The state-of-the-art multimodal model Self Supervised Embedding Feature Transformer (SSE-FT) was finetuned and assessed on the Multimodal EmotionLines Dataset (MELD), revealing performance issues. The interpretability framework MM-SHAP was modified for emotion recognition and extended to include the text, audio, and video modalities. The proposed interpretability framework and ablation studies showed the SSE-FT predominately relied on the textual modality, leading to uni-modal collapse. The Dutch language model RobBERT was integrated into SSE-FT to increase performance, yet training RobBERT independently showed its limitations in capturing nuanced emotional cues from the MELD dataset. This thesis introduces visualization techniques specifically developed to focus on increasing interpretability within individual modalities, and to assist comparative analysis between the audio and text modality. The proposed interpretability method and visualization technique for text is applied to analyze the textual modality and show valuable insights into the model's learned emotional cues for the textual modality. The results show that SSE-FT trained on MELD relies heavily on paralinguistic cues in text and is not able to capture the more nuanced emotional cues in the video and audio modality. The findings of this thesis call attention to the need for a balanced, high-quality Dutch dataset for emotion recognition as well as the importance of general dataset quality for advancing in the field. The proposed interpretability method is found to be effective for creating interpretability in multimodal models for emotion recognition. **Keywords:** Multimodal, Emotion Recognition, Interpretability, SSE-FT, MM-SHAP, Uni-modal Collapse, Visualization

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

ASR Automatic Speech Transcription

BERT Bidirectional Encoder Representations from Transformers

CNN Convolutional Neural Network

FER Facial Emotion Recognition

IAA Inter Annotator Agreement

LSTM Long Short-Term Memory

MELD Multimodal EmotionLines Datase

MFCC Mel-Frequency Cepstral Coefficients

MMER Multimodal Emotion Recognition

MM-SHAP Multimodal Shapley Additive Explanations

NLP Natural Language Processing

SER Speech Emotion Recognition

SSE-FT Self Supervised Embedding Feature Transformer

SVM Support Vector Machine

Chapter 1

Introduction

The introduction section provides the motivation, goals and scope for this research. To begin, the background and context for the research are given in Section 1.1 and 1.2. Following this, the main research question and subquestions are outlined in Section 1.3. Furthermore, the relevance of the current research within the field of Artificial Intelligence (AI) is explained in Section 1.4. Lastly, the structure of the current thesis is given in Section 1.5.

1.1 Preface

Large data collections are treasure chests for new research. However, the volume and complexity of these datasets pose challenges for extracting useful information. The Dutch Institute of Sound & Vision [1] is such a large data collection, as it manages one of the world's most extensive media archives, including radio shows, TV shows, YouTube videos, written press, podcasts, and games.

Over decennia, the media in the archive has captured influences on Dutch culture and our societal development. In recent years, the media landscape has changed vastly with the explosion of social media, which now plays a serious role in our everyday lives and in shaping our modern culture. 'Mass media', or "the communication (written, broadcast, or spoken) that reaches a large audience", influences how we think, behave, and perceive others and the world around us [107]. As the world around us changes, it is even more important to keep up to date with the influences of media on our society, and develop the appropriate tools to be in charge of the transformation.

The aim of the Institute of Sound & Vision is to preserve the media in a sustainable manner and explore the potential use of the archive. To achieve this, the institute supports research on the archive and makes efforts to improve the archive's accessibility.

[1] https://beeldengeluid.nl/

To inspire researchers to study on the archive, they create building blocks that can be used for research, such as segmentation, speaker diarization, and feature extraction systems. To describe the process of research on the archive, data stories are created to tell a story on a specific topic of interest based on analysis of data from the archive.

The initial collaboration for this thesis stems from the idea of applying machine learning techniques to extract nuanced sentiment and/or emotions from media in the Sound & Vision archive. Because the archive is inherently multimodal, including many types of videos such as talk-shows, interviews, or documentaries, the idea came to go beyond emotion recognition from text and exploit the visual and audio modality to capture more nuances in emotions from the archive. Namely, performance on the task of emotion recognition, is strengthened by including multiple modalities [17, 61, 86].

Hence, it was decided to explore the deployment of a multimodal model for emotion recognition to be used as a building block for research on the archive. To verify the reliability of this fairly complex multimodal model for emotion recognition, a layer of interpretability is added to warrant the validity of the academic research for which this system is used.

Moreover, the proposed interpretability approach provides insights into the emotional cues used by the model in each modality. The outcomes of the current thesis will contribute to the stimulation of journalistic and academic research and, thereby, our understanding of Dutch cultural and societal phenomena hidden in the large media archive of Sound & Vision. The interpretable multimodal emotion recognition system is implemented in the pipeline used to work with data from the archive, called 'Dane', to ensure proper application functioning.

1.2 Background

Due to the fast technological advances in machine learning and the increase in computing power, computers are nowadays able to perform well in human tasks, including more intricate tasks such as emotion recognition. Emotion recognition systems are widely used in applications such as mental health monitoring and customer satisfaction [12, 7]. In such applications, it is beneficial for these computer systems to understand emotional states to effectively reply to users' needs. Moreover, in video streaming platform interfaces, emotion recognition systems could be used to improve personalized recommendations from the emotional content in a video[5].

It is hard to break down human emotions, but an automatic system can, for instance, recognize them from our facial expressions, speech, and behavior. Recognizing emotions, however, remains a challenging task for the following reasons: Emotions are dependent on individual, cultural, and societal factors. The way emotions are expressed and perceived can vary a lot between one individual and another. We all have different

body language, facial properties,and uses of voice and language. Moreover, emotions can vary in different situational contexts. Small nuances can make the difference between a person using irony or being completely serious. Furthermore, difficulties in the annotation process of emotions arise from these nuances and individual differences as there is no clear set of rules on how to annotate emotions [43].

The features used in emotion recognition systems are tied to how humans express emotions and extend across different information modes or modalities, such as text, audio, or video. Multiple emotion recognition systems using information from only one of those modalities have been proposed, demonstrating efficiency in specific contexts. These so-called 'uni-modal emotion recognition systems', traditionally select handcrafted features such as selecting emotional words [50] and crafting acoustic features [82]. Nowadays, more sophisticated approaches have been proposed using automatic feature extraction methods and machine learning methods [46, 33, 76, 45].

Emotion recognition, although applicable within one modality, can benefit from the integration of multiple modalities [17, 61, 86]. Some machine-learning tasks, such as image captioning and visual question answering, rely on multiple input sources or modalities. These tasks are inherently multimodal. The use of multiple modalities fits, in general, within the current developments of creating more general and robust AI and is possible due to the rapid advances in technologies that allow processing multiple inputs in parallel. Numerous multimodal models for emotion recognition have been introduced.

To handle the multimodal data, these models use different fusion strategies. In early fusion, modalities are combined at the feature level before predictions are made. On the contrary, late fusion approaches involve a separate process for each modality and the fusing of their predictions, often using mechanisms such as voting [9]. Moreover, diverse hybrid fusion methods, which combine both early and late strategies, have been introduced in the literature [115, 80, 44, 1, 104]. Deep learning models are complex, and they are often described as 'black boxes' due to the challenge of understanding precisely how they arrive at their predictions. The field of Explainable Artificial Intelligence (XAI) aims to improve the interpretability of these models, making their predictions more understandable and reliable [49].

The complexity of machine learning models increases when multiple data types are added. As modalities have interactions and dependencies with each other, the model must learn how different modalities influence and relate to one another. Multimodal models trained on large datasets might lack a real understanding of the different modalities, relying more on learned statistical patterns. Moreover, multimodal models might reduce themselves to a uni-modal model, misusing biases in one modality or relying on one modality only. These models are not able to use the information from the other modalities, resulting in a so-called 'uni-modal collapse' [68]. Hence, the need for interpretable methods that assess the multimodality of these models and make their

decision-making process transparent, becomes apparent.

Multiple frameworks for adding interpretability have been proposed. SHAP and LIME are model-agnostic methods, meaning that they can be applied to various models without being dependent on the specifics of any particular model architecture [65, 84]. Moreover, within the transformer architecture, the attention mechanism can be leveraged to improve interpretability [25].

These interpretability frameworks can be extended to involve more modalities, as Parcalabescu et al. did through the introduction of MM-SHAP (Multi-Modal SHAP) [75]. The authors introduce MM-SHAP as a modality contribution metric, focusing on Image-text models while excluding the audio and video modalities. Evaluation is limited to Visual Question Answering and Image-sentence alignment tasks. As the number of multimodal models for emotion recognition grows, the need arises for an interpretability method that is tailored for this task. Extending MM-SHAP to the audio and video modalities would increase its use by assessing a larger range of multimodal models. Moreover, extending the method to also be able to provide interpretability within the modalities would provide a more detailed assessment of the robustness of a multimodal model.

1.3 Research objectives

The aims of the current research are the following: In collaboration with the Dutch Institute of Sound & Vision (S&V), the deployment of an interpretable multimodal model is explored. The state-of-the-art multimodal model 'Self Supervised Embedding Feature Transformer' (SSE-FT) is implemented for the S&V pipeline to be used as a building block for research on the archive.

To increase transparency and reliability in the implemented multimodal model, an interpretability framework for emotion recognition is proposed and implemented for SSE-FT. The MM-SHAP method for assessing multimodality is extended for video and audio and modified to provide interpretability within these modalities. Visualizations are created to manage a straightforward analysis of the results.

SSE-FT will be finetuned and evaluated on the 'Multimodal EmotionLines Dataset' (MELD). MELD is a multi-party dataset for multimodal emotion recognition with clips extracted from the TV-series 'Friends' [81]. The proposed interpretability framework can be used to further analyze the global performance and local predictions of SSE-FT on the MELD dataset. Modality contributions of audio, video, and text on different levels (sample, label, and dataset) can be investigated. Furthermore, the framework can help spot the emotional cues used by the model within the modalities, which could detect possible biases in the model.

Implementing SSE-FT with an interpretability framework on the S&V archive func-

tions as a good use case for exploring interpretability in multimodal emotion recognition. This leads to the central research question of the current thesis: How can we improve the interpretability of multimodal models for emotion recognition with the use of MM-SHAP?

Sub-questions supporting this research question can be posed to guide the research:

1. How well does the multimodal transformer SSE-FT perform at emotion recognition using the MELD dataset?
2. How well does SSE-FT perform at recognizing emotions from videos in the Sound & Vision archive?
3. How can the MM-SHAP method be extended to incorporate the video and audio modalities?
4. How can the MM-SHAP method be used for emotion recognition?
5. How can the MM-SHAP method be extended to provide interpretability within modalities?
6. What is the modality contribution of SSE-FT at sample, emotion label, and dataset level on the MELD dataset according to the interpretability framework?
7. How can the interpretability framework be used to discover emotional cues that SSE-FT has learned from the data?
8. How can visualizations be created and used to increase interpretability within modalities?
9. How can visualizations be created and used to analyze certain emotional properties within modalities?

1.4 Contribution

The current research holds significant relevance in the field of AI, as the main goal is to improve the interpretability of state-of-the-art multimodal transformer models for emotion recognition. Most research on interpretability within multimodal models focuses on two modalities, text and image. The model implemented in the current research uses text, audio, and video, as these modalities provide important emotional cues.

The current research investigates the finetuning process on the benchmark dataset MELD and the decision-making process of SSE-FT on these samples. The MELD dataset is the only multi-party dataset for emotion recognition, and performance on

the dataset is in general quite low. The current research has detailed answers to the behavior of SSE-FT on the MELD dataset.

The proposed methods provide exhaustive analysis and interpretability on SSE-FTs decision making process, as both global and local interpretability is given. Designs for visualizations are presented, visualizing how different parts of each input modality influence the prediction of a sample. This contributes to the development of more transparent and reliable AI systems.

Furthermore, by collaborating with the Institute of Sound & Vision, the results from this research are directly evaluated on a real-world application. The implemented interpretable emotion recognition system provides a reliable base for future research. The current research offers grounded recommendations for future development, contributing to research on extracting emotion from media archives. Moreover, the results from this research will hopefully inspire the creation of a Dutch multimodal dataset for emotion recognition.

1.5 Thesis structure

First, a review of the related work and topics related to the research of the current thesis is given in Chapter 2. The concepts and related work regarding natural language processing, emotion recognition, multimodal emotion recognition, datasets for multimodal emotion recognition, and interpretability methods are explained in Sections 2.1, 2.2, 2.3, 2.4, and 2.5.

In the methodology in Chapter 3, an overview of the research is given in Section 3.1, the dataset used in the current research is described in Section 3.2, the implementation details of SSE-FT are given in Section 3.3, the proposed interpretability framework is described in Section 3.4, the visualization designs are presented in Section 3.5, and details regarding the implementation for the Institute of Sound & Vision are given in Section 3.6.

The results regarding the performance of SSE-FT on the MELD dataset (Section 4.1) and the interpretability framework (Section 4.2) will be presented in Chapter 4.

In Chapter 5, the performance of SSE-FT is analyzed in Section 5.3, the implementation and results from the interpretability framework are discussed in Section 5.2, the performance of SSE-FT on the S&V archive is discussed in Section 5.3, limitations of the current research are described in Section 5.4 and recommendations for future work are given in Section 5.5.

In the conclusion in Chapter 6, a summary of the findings of this research is given.

Chapter 2

Related work

This chapter starts with a review of the existing literature on interpretable multimodal models for emotion recognition. First, the topics of natural language processing and emotion recognition are explored in Section 2.1, and Section 2.2. Subsequently, multimodal emotion recognition (MMER) is addressed in Section 2.3 and datasets frequently employed for this task in Section 2.4. The chapter concludes with Section 2.5, with an emphasis on the significance of interpretability within multimodal models for emotion recognition and an overview of available interpretability methods.

2.1 Natural language processing

Natural Language Processing (NLP) is the field of research that focuses on how computers can be used to understand and generate text or spoken language to perform tasks [50, p.60]. Researchers in the field of NLP have the goal of gathering knowledge on how humans understand and use language, intending to create methods to make computer systems understand and manipulate natural languages to execute specific tasks [29]. Within the domain of NLP, a great number of valuable tasks exist, including part of speech tagging (PoS-tagging), information retrieval (IR), named entity recognition (NER), language translation, text generation, question answering (QA), and sentiment analysis (SA).

2.1.1 Multimodal models

Machine learning models specialize in processing a single type of input and generating a corresponding type of output [50]. For example, models can be designed to perform tasks such as translating one language to another or predicting new numerical values from a series of other numerical values. Input sources of these models can consist of various modalities such as text, audio, images, and videos.

When combining modalities in a machine learning model, and creating a multimodal model, these models can capture additional dependencies within the data, leading to a better understanding of the information [9]. In addition to improving performance and ensuring more robust predictions, multimodal models are capable of functioning effectively even in scenarios where some modalities contain little to no information. In such cases, another modality can compensate and contribute to the task performance [9]. Furthermore, machine learning models will, when combining multiple modalities, be closer to operating in a manner that closely resembles human cognitive processing. Human decision-making often relies on multiple information sources as well.

Multimodal models have a wide application, such as but not limited to education [8], autonomous vehicles research [117], and the healthcare sector [7]. In NLP, the motivation for incorporating additional data input sources stems partly from the challenges posed by lexical ambiguity and encountering out-of-vocabulary words [40]. As a solution for this problem, other modalities can provide more context or clues for disambiguation. For instance, consider the sentence 'He swung the bat'. The word 'bat' can have two meanings: an animal or an object used in baseball. When the sentence is accompanied by an image that shows a man swinging a baseball bat, the visual context provided by the image clarifies that in this specific instance, 'bat' refers to the object used in baseball. Hence, the sentence is disambiguated.

Moreover, NLP tasks can benefit from the extra information gained by adding more modalities. Such tasks include sentiment analysis and emotion recognition. Finally, some NLP tasks inherently involve the integration of information from different modalities, e.g., image captioning (IC), visual question answering (VQA), and speech-image-text alignment (SITA). Even though in theory including multiple modalities has benefits, it also presents a range of challenges, as described in Section 2.3.

2.2 Emotion recognition

Emotion recognition and sentiment analysis, two important branches of NLP, traditionally study the feelings and emotional nuances embedded in textual and spoken language. Sentiment analysis focuses on the classification of a text into either a positive, negative, or neutral tone, often using sentiment scores to capture the full range of emotional expression. Sentiment can be detected, for example, at the document level or in a dialogue. Moreover, stance detection focuses on detecting sentiment directed toward specific subjects or individuals [56].

Emotion recognition goes beyond this more simplistic classification, and aims to understand a broader aspect of human emotions. In Section 2.4.1.1, emotions are categorized in a dimensional model, where emotion recognition incorporates both valence and arousal. Valence is similar to sentiment as it indicates how positive or negative

an emotion is. Another dimension, arousal, is added to capture the intensity of the emotion [85].

Human emotions can be recognized from facial expressions, speech, and behavior. Humans, however, can conceal their emotions from external observation. For this reason, recent studies have been focusing on physiological signals such as recordings of brain activity (EEG), since these signals provide a more genuine representation of a person's emotional state [98]. In the next sections, Section 2.2.1 discusses how emotional features are extracted, and Section 2.2.2 provides information on different models for emotion recognition.

2.2.1 Feature extraction for emotion recognition

To recognize emotions, it is first necessary to extract the emotional states from the data by transforming the data into informative representations of the data. This process is what is known as feature extraction [50]. In the next sections, feature extraction techniques for text (Section 2.2.1.1), audio (Section 2.2.1.2), and visual (Section 2.2.1.3) information sources will be explained.

2.2.1.1 Textual features

There are multiple feature extraction methods for text. A popular NLP technique is 'Bag-of-words' (BoW), This technique represents a piece of text as a collection of individual words without keeping the order and structure of the words. It essentially creates a 'bag' of all the words in a document, together with their frequency of occurrence. Since certain words' presence often indicates the emotional tone of a text, BoW can be a helpful technique for emotion recognition [50].

The 'Term Frequency-Inverse Document Frequency' (TF-IDF), takes into account both the frequency of a word and the rarity of the word across documents. The method indicates if a word is important within a specific document but not overly common in the entire corpus [88]. TF-IDF is often used to find the words most important for a certain task, such as emotion recognition. What words contribute most to the emotional load of the text?

'Word embeddings', project words into a dense vector space, where words with similar meanings are positioned close to each other. This method captures semantic relationships between words. However, the meaning of a word can be context-dependent. While static word embeddings such as GloVe [77] and Word2Vec [72] maintain a consistent embedding for a word regardless of its context, contextual embeddings such as BERT [34] and other attention-based methods change the words' embedding based on the surrounding context. These methods allow the same word to have different embeddings based on how it is used in context.

2.2.1.2 Audio features

For speech emotion recognition (SER), features such as pitch, intensity, and prosody can be extracted. The pitch of a sound is the perceived fundamental frequency, which corresponds to how high or low a person's voice sounds. Intensity measures the amplitude of an audio signal, which corresponds to the loudness of a sound. Prosody relates to the rhythm, timing, and intonation of speech. Prosody includes multiple features such as speech rate, the total length of a speech, pauses, and intonation patterns [82]. These features capture variations in speech patterns, which can indicate (changes in) emotional states [86]. For example, a higher pitch might indicate the emotional state of happiness, while a lower pitch might indicate anger. A rapid speech with frequent pauses might indicate nervousness, while a steady and slower pace could suggest calmness or sadness [82].

While prosodic features provide valuable insights into emotion recognition, they are often insufficient when used alone in speaker-independent algorithms. To increase performance, phonetic features like Mel-frequency cepstral coefficients (MFCC) are incorporated [91]. MFCC focuses on the smallest units of speech (phonemes), while prosodic features consider larger speech segments. MFCC captures the spectral characteristics of audio by representing the short-term power spectrum of a signal. MFCC is shown to be an effective feature for speech emotion recognition [91].

Moreover, recent advancements in SER have introduced end-to-end approaches that bypass the need for specific features like pitch, intensity, or MFCC's. These approaches directly encode the waveform data with neural networks such as Wav2Vec2. Researchers have demonstrated the effectiveness of such end-to-end methods [46, 33, 76, 45]. Furthermore, researchers have developed specialized toolboxes for audio feature extraction, such as OpenSMILE [39] and COVAREP [30]. OpenSMILE includes features such as those mentioned before and is made for real-time operations because it is simple and fast. Nowadays, it is widely used in the field of SER. COVAREP is a specialized toolbox that focuses mainly on the emotional aspects of speech, it extracts acoustic features that are highly relevant to emotion analysis and SER as well.

2.2.1.3 Visual features

Visual features for emotion recognition focus on facial and bodily gestures. 'Facial landmarks', are points on the face, including the eyes, nose, and mouth, that give away emotional cues. Tracking these landmarks can tell us about a person's emotional state [38]. Moreover, combinations of certain muscles in our face produce emotional expressions. These specific facial muscles are defined by the Facial Action Coding System (FACS) [38]. Different combinations of these muscles, known as 'AUs', produce various emotional expressions. For example, a combination of action units 6 (cheek

raiser) and 12 (lip corner puller) signifies happiness. Automatic systems such as FACET (Facial Action Coding System) and OpenFace [10] use features based on FACS, along with techniques such as eye gaze tracking and microexpression analysis.

2.2.2 Models for emotion recognition

Models for emotion recognition have been developed across various data sources, e.g., text, audio, and images, and significant advancements have taken place. In the next sections, textual (Section 2.2.2.1), audio (Section 2.2.2.2), and visual (Section 2.2.2.3) models for emotion recognition are discussed.

2.2.2.1 Textual models

In the last decade, emotion recognition through text has witnessed a transformation. Traditionally, text-based emotion recognition involved selecting emotional keywords, incorporating BoW representations, and employing N-grams. However, these methods often struggle with sentences where emotional keywords might not be explicitly present as the data might be sparse [92].

Researchers have introduced more sophisticated approaches, focusing on textual features and improving machine learning models. An approach by Alm et al. involves supervised machine learning and achieves high accuracy with a broad range of textual features [4]. Additionally, Liu et al. leverage a real-world knowledge base known as Open Mind, which contains a repository of 400,000 pieces of knowledge [62]. Moreover, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a highly effective tool [34]. In Section 2.3.4.1, there will be a breakdown of this model.

2.2.2.2 Speech models

Earlier, SER relied on acoustic features and machine learning algorithms. An approach by Milton et al. combines MFCC features and feature engineering with a traditional machine learning technique, Support Vector Machines (SVM), to categorize emotions [73].

In more recent developments, a hybrid CNN-LSTM deep network has been specially designed for audio emotion classification, including both speech and song descriptions. The authors used MFCC for feature extraction and achieved 73.33% accuracy for emotions extracted from audio songs and 53.32% accuracy for emotions extracted from audio speech [6]. Another novel approach involves a deep graph method to address the task of SER. The authors represent the speech data in the form of graphs and achieve competing performance using fewer parameters than other SER models [96]. Furthermore, in a study by Gong et al., a transformer architecture was created for SER. It is

common to add a self-attention layer to a CNN to improve focus on global dependencies. However, the authors achieve great performance using self-attention without the CNN architecture [45].

2.2.2.3 Visual models

As previously discussed in Section 2.2.1.3, visual models for emotion recognition often use features based on facial attributes. In the domain of Facial Expression Recognition (FER), two main approaches exist: frame-based FER and video-based FER. The first approach, frame-based FER, uses static facial features from images or selected peak expression frames from image series. The second approach, video-based FER, leverages spatio-temporal features to capture the dynamic changes in facial expressions over time [52]. In many video-based FER models, the system tracks the facial landmarks over time, frame by frame. This tracking captures not only the spatial characteristics of facial expressions but also introduces an additional dimension: time. For instance, one approach by Ghimire et al. employs facial landmark displacement as a feature, extracted with the AdaBoost algorithm, and applies an SVM for expression classification [42]. Approaches like these rely on handcrafted features for extraction, followed by a (pre-trained) classifier, such as a 'Suport Vector Machine' (SVM) or 'Random Forest', for the classification task. These classifiers are known for their time efficiency and low computational resource requirements, which makes them good alternatives to deep learning approaches.

In recent times, Convolutional Neural Networks (CNNs), a deep learning method, have been effective in various visual machine-learning tasks, including FER. In this deep learning approach, input images undergo convolution operations to create feature maps. These feature maps are then combined with fully connected networks to classify the facial expression into emotional categories [3]. The usefulness of CNNs is shown in a study in which a simple CNN was implemented for the experiments. The authors visualize the features of the CNN to understand the feature maps that are obtained by training [13]. They discover a correlation between the features generated by the unsupervised learning process and Ekman's AU's. Nonetheless, CNN-based methods, while effective for spatial features in individual frames, may not fully capture the temporal variations in facial components. To address this limitation, a recent hybrid approach combines a CNN to extract spatial features from individual frames and incorporates Long Short-Term Memory (LSTM) to capture temporal features across sequential frames [52]. Although deep learning approaches achieve great performance in FER, these models need a significant amount of training data and computational resources (depending on the specific model).

2.3 Multimodal emotion recognition

Emotions are complex, and understanding them thoroughly often requires more than just a single source of information. Research findings show that emotion recognition benefits significantly from the integration of speech, vision, and text information [17, 61, 86]. In the following sections, there will be an explanation of multimodal emotion recognition. The following aspects will be discussed as they are relevant to multimodal models: representation (Section 2.3.1), alignment (Section 2.3.2), and fusion methods (Section 2.1).

2.3.1 Representation

One of the fundamental challenges in multimodal learning is effectively representing and choosing features for data from multiple modalities. This is challenging because each modality presents the data in a different form, ranging from textual representation as BoW or word embeddings to images represented by pixel values or deep feature representations, and audio represented as raw waveforms or MFCC's.

Moreover, each modality conveys a different form of information [9]. The textual modality gives semantic and symbolic information through language. The visual modality provides aspects like color, shapes, and spatial relationships. Audio information includes sound frequencies and patterns. The way these modalities relate to emotions is also different. Words and sentences can directly imply emotional states, while visual and audio convey more indirect nonverbal cues such as facial gestures or the pitch of a voice.

Furthermore, the dimensionality of the modalities can diverge. Text can be considered one-dimensional in terms of spatial information, as it lacks spatial variations. In contrast, audio is inherently one-dimensional in a spatial sense, but it has an additional dimension to represent temporal information, essentially making it two-dimensional to account for time. The presence of a temporal dimension is inherent to the concept of a video itself, whether the video is represented in a sequence of individual frames or as a continuous stream of differences between sequential frames [9].

An alternative approach to representation involves using uni-modal features (features including only a single modality), which can be crafted by hand or generated through deep learning techniques as discussed in Section 2.2.1. Nowadays, the latter approach is most commonly practiced. To use a neural network for data representation, the network is first trained for a task, for example, recognizing words from speech. Deep neural networks have multiple layers, and each successive layer represents the data more abstractly. The last, or one of the late neural layers, is therefore often used to represent the data[9].

Various other sophisticated techniques can be used for multimodal representation,

including Probabilistic Graphical Models (PGM), sequential models such as Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks, and more recently, transformers. PGMs use random latent variables to capture hidden patterns in the data. These probabilistic models can describe how different modalities contribute to the total likelihood of observed data. One particular approach involves constructing 'Multilayer Boltzmann Machines' and shows that these systems can be used in well-performing generative models [87]. Probabilistic graphical models handle incomplete or missing data by making reliable estimates and can uncover patterns in unlabeled data by using the underlying probability distributions[87]. One notable limitation is the computational cost associated with training these models.

Furthermore, sequential models such as RNNs and LSTMs are widely used for modeling sequences in multimodal data, demonstrating success in emotion recognition, and achieving good performance compared to baseline methods [28]. Transformers, in contrast to RNNs and LSTMs, process data without relying on sequential processing and have gained popularity for multimodal tasks due to their ability to capture dependencies among elements in a sequence, regardless of their relative positions. The architecture of transformers and their use in (interpretable) multimodal models is elaborated in Section 2.3.4.

2.3.2 Alignment

Multimodal alignment can be defined as "finding relationships and correspondences between sub-components of instances from two or more modalities" [9]. For example, the alignment process aims to find the exact correspondence between textual subtitles and audio content. Alignment methods are mostly based on finding similarities between segments from uni-modal data.

Early research focused on aligning multimodal sequences unsupervised, with methods such as dynamic programming and generative graphical models. For instance, dynamic time warping is used as an approach to assess the similarity between two sequences to find their best alignment [55]. Similarly, graphical models are constructed in a study by Yu et al. to align visual objects in images with spoken words [114].

These methods aim to establish correspondences between multimodal data without the need for labeled annotations. However, with the growth of datasets with annotated labeled instances, supervised alignment methods have gained popularity. These approaches are often based on deep learning [69, 78]. The Automatic Speech Recognition (ASR) model 'Whisper', for example, uses supervised learning to extract text from speech [83]. Whisper provides timestamps that can, for instance, be used to align the text with audio and video.

In the current thesis, a dataset with pre-aligned multimodal data is used, the dataset includes timestamps in the audio and videos. Such datasets, in which alignments are

explicitly annotated, are valuable for training and evaluating multimodal models. However, datasets with explicitly annotated alignments are sparse and often do not fit the desirable aligned format. Such datasets have to be aligned manually or via one of the automatic approaches mentioned in this section.

2.3.3 Fusion methods

Fusion methods are an important aspect of multimodal learning because they determine how information from different modalities is combined for effective modeling. These methods can be classified as early fusion and late fusion. Moreover, hybrid fusion methods combine the strengths of both methods to increase performance. Often, early and late techniques make use of neural networks discussed in Section 2.1.

More recent fusion methods are designed to be a better fit for multimodal data [41]. Such methods include multiple kernel learning, graphical models, and neural networks. In the scope of this research, early and late fusion methods are discussed (Section 2.3.3.1 and 2.3.3.2), as well as fusion methods using neural networks (Section 2.3.3.3).

2.3.3.1 Early Fusion

Early fusion, also known as feature-level fusion, involves extracting features from each modality, integrating them, and feeding the combined features to a classification model. The simplest early fusion method is concatenating the features into one input vector. More advanced techniques involve creating a joint representation vector using neural networks. The topic of joint representation is further elaborated in Section 2.3.3.3.

Early fusion allows the exploration of interactions between raw features across modalities. However, this strategy also presents a challenge: as features from different modalities often represent different physical properties, and the classifier must learn both feature abstractions and their interactions simultaneously. This can lead to high-dimensional input spaces and potential computational complexity, making the model prone to overfitting [101]. A schematic overview of the early fusion approach can be seen in Figure 2.1.

2.3.3.2 Late Fusion

Late fusion techniques focus on training an uni-modal model for each modality separately and then fusing their predictions. Fusion can be achieved through methods like voting, weighing, or training an additional model, such as a neural network, to combine the predictions. Late fusion allows each modality to be processed independently, using the strengths of individual modality-specific models. However, this approach can, in some cases, ignore low-level interactions between modalities [110]. Despite this limitation, the late fusion strategy has been used successfully in various applications,

achieving competitive performance. An approach by Tripathi et al. incorporates late fusion and obtains high performance on the IEMOCAP dataset [103]. Another approach uses late fusion for MMER that relies on speech and facial information and achieves high accuracy on the RAVDESS dataset [64]. A schematic overview of the late fusion approach can be seen in Figure 2.1.

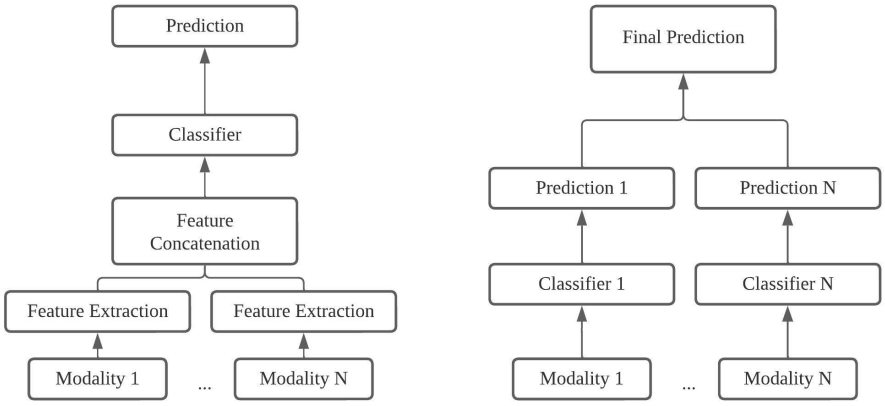


Figure 2.1: A schematic overview the early fusion method (left) and the late fusion method (right) for multimodal models.

2.3.3.3 Fusion with neural networks

Both early and late fusion methods can make use of a neural network model to fuse the modalities. In this approach, a neural network maps individual modalities onto a unified shared representation vector space, either via joint representation or coordinated representation. [49, 9]. An example of an early fusion strategy with neural networks involves the use of fine-tuned neural networks to generate embeddings from text and audio [32]. The embeddings are concatenated and fed into a transformer model equipped with co-attention mechanisms, using the most important parts of each embedding. This process results in a joint representation, which is then used as input for the classifier. This technique achieves high performance on the IEMOCAP and SAVEEE datasets for emotion recognition. Moreover, a late fusion technique is implemented by Sun et al., where audio, text, and images are separately processed using a Bi-LSTM with an additional self-attention layer. The outputs of these modality-specific models are then fed into another Bi-LSTM for the final prediction [101].

Additionally, many deep learning approaches employ a hybrid fusion technique. In one approach by Zadeh et al., a tensor fusion network is introduced to express mul-

timodal fusion information through image, audio, and visual features [115]. In another method, an LSTM is applied separately to text, visual, and audio data, and the extracted features are integrated into a multi-level fusion learning architecture [80]. Ghosal et al. propose a multi-attention RNN framework to learn features using attention for multimodal representation [44].

2.3.4 Transformers

The transformer architecture was introduced in the paper 'Attention is all you need' [108]. The architecture of the model is that of an encoder-decoder. First, all input is processed and made into embedding vectors. These embedding vectors are put into the encoder, which consists of three components; positional encoders, multi-head attention, and a feed-forward layer. Positional encoders are used to handle the meaning of the words in different sentences based on their position in the sentence.

Attention is used to give context to the numerical vectors representing the words. The attention layer determines what part of the input the focus should be on. The attention mechanism can be multi-headed, meaning it computes multiple attention vectors and averages them per word, capturing cross-word relations. Because of the attention mechanism, the input data can be handled in parallel, thus eliminating the need for sequential processing.

The feed-forward layer simplifies the information obtained from attention, by reducing the dimensionality of the data and applying non-linear transformations. Each attention vector is processed by its own feed-forward layer, making it a parallel and fast process.

Transformers are often used for sequence-to-sequence tasks, such as machine translation, and include decoder layers for this purpose [108]. However, for tasks like emotion recognition, which do not require generating new sequences or translations, encoder layers, in combination with classification layers and an activation function, form the appropriate components for the task.

2.3.4.1 BERT

BERT, or 'Bidirectional Encoder Representations from Transformers', is a transformative deep learning model that has had a significant impact on NLP and many other domains [34]. BERT can accurately generate and understand human language, since the model is trained on a large corpus of textual data. The architecture of BERT makes use of a 'Masked Language Model', which masks tokens in the input text randomly, forcing the model to make predictions about the original token identities based on their context. BERT is widely used in multimodal models as a text-encoder, and moreover, specialized variants of BERT are developed to incorporate visual and audio data.

2.3.4.2 Visual transformers

'Visual Transformers' or ViT, draw inspiration from the NLP transformer architecture previously described in Section 2.3.4 and adapt this to images or video. The Transformer's self-attention mechanism is used to capture dependencies over long distances and contextual information within the visual data. The images are broken up into patches and then flattened, as can be seen in Figure 2.2. Moreover, Figure 2.2 shows the ViT architecture in comparison to the original Transformer encoder architecture [35]. ViTs are powerful in tasks such as image classification [35], object detection [20], semantic segmentation [113], and emotion recognition [23] [67].

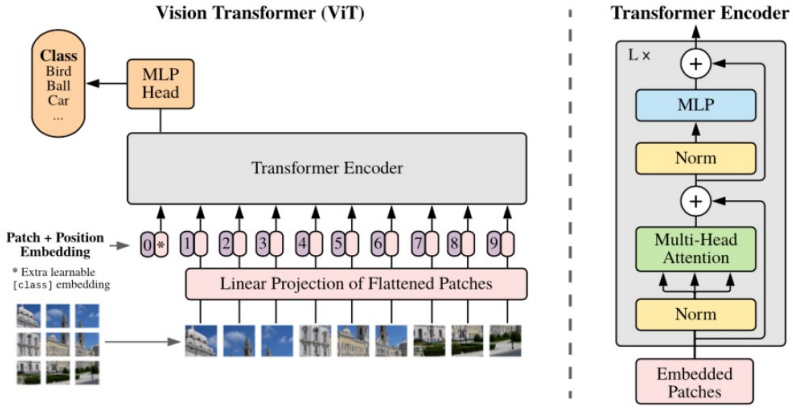


Figure 2.2: The architecture of the ViT (left) and the architecture of the original Transformer encoder (right), Figure from [35].

2.3.5 Multimodal transformers

Multimodal transformers represent a significant advancement in deep learning. These models can combine the strengths of different uni-modal models, such as BERT and ViT, to simultaneously process diverse modalities. ClipBert and Video-BERT are two noteworthy transformers designed for video-text understanding. ClipBert has a single-stream early fusion architecture, in which the two modalities are first fused and then processed together [59]. In the training phase, ClipBert uses sampled short clips at each training step, which makes the model efficient for a small training dataset. It uses a 2D CNN architecture for the video encoding: ResNet-50. Video-BERT, on the contrary, has a dual-stream late-fusion architecture where the two modalities are processed apart and later fused [100].

Transformers, processing three modalities (video, audio, and text), have been pro-

posed as well. MEmoBERT is a self-supervised model for MMER [121]. This model learns joint representations across modalities through self-supervised learning on an unlabeled video dataset. MSAF, or 'Multimodal Split Attention Fusion', is a multimodal transformer for emotion recognition. It makes use of a special type of fusion technique. Each modality is divided into channel-wise equal feature blocks, and a joint representation is formed to produce soft attention for each channel across the feature blocks [99]. HERO, or 'Hierarchical Transformer Architectures' has a hierarchical architecture, consisting of a cross-modal transformer and a temporal transformer for multimodal fusion [60]. SWAFN, or 'Sentiment Words Attention Fusion Network', is a model designed for SA, trained on the CMU-MOSI and CMU-MOSEI datasets [27].

In the next sections, three multimodal transformer models are explored in depth. As each of these models accepts raw input, they are suitable for the archive pipeline of Sound & Vision. After attempting to run both MulT and VATT models, it appeared that they were not reproducible. Therefore, the multimodal transformer SSE-FT will be used in the experiments. The motivation behind choosing this model will be further elaborated on in Section 3.2.

2.3.5.1 MulT

In the MulT architecture [104], depicted in Figure 2.3, first the input from each modality goes through a convolutional layer to extract the local structure of the input. Next, positional embeddings are added to enable the input to carry temporal information. Next, the input goes through cross-modal transformers, so each modality can use information from the other modalities. In a cross-modal transformer, the target modality is repeatedly reinforced with low-level features from another modality. This is done by learning attention scores across the features of the two modalities. The authors suggest that adapting from low-level features is advantageous for the model because it helps preserve important low-level information specific to each modality. Finally, the outputs are concatenated from the cross-modal transformers that have the same target modality and passed through a self-attention transformer. The final elements of these self-attention transformers are extracted and used for prediction.

2.3.5.2 VATT

The design of VATT [1], and its self-supervised learning strategy are depicted in Figure 2.4. The model takes raw video, audio, and text. There are two primary configurations in the VATT architecture. In one, each modality has its own transformer with specific weights. In the second configuration, all modalities share a single transformer, meaning one transformer is applied universally across all modalities.

The architecture involves a tokenization layer, embedding vectors, and transformer

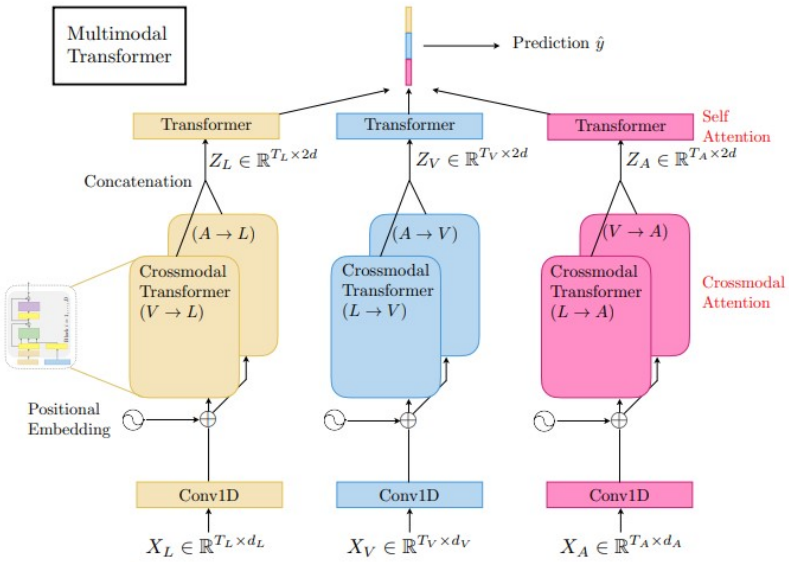


Figure 2.3: The architecture of MulT for modalities (text (L), video (V), and audio (A)), crossmodal transformers serve as the core components for the multimodal fusion [104].

components. VATT first converts each modality into a feature vector through linear projection before inputting it into a transformer encoder. Each modality is processed with its own positional encoding, so the transformer can distinguish between tokens based on their position in the input sequence. VATT makes use of DropToken, which is a mechanism that randomly selects certain input tokens to be processed, reducing the computational cost of the transformer. The transformer architecture includes a multi-head-attention module that employs a standard self-attention mechanism. This allows the model to attend to different positions in the input sequence simultaneously, as previously discussed in Section 2.3.4. The activation function used in the multi-layer perceptron layer is the Gaussian Error Linear Unit (GeLU), chosen for its effectiveness in capturing complex relationships within the data. Moreover, in the multimodal projection head, a semantically hierarchical common space mapping is created to compare embedding pairs of video-audio as well as video-text with their cosine similarity. This is done to take the semantic granularity of these modalities into account. Furthermore, alignment of video-audio embedding pairs is done with Noise Contrastive Estimation (NCE) and Multiple Instance Learning NCE (MIL-NCE) is used to align video-text embedding pairs [2].

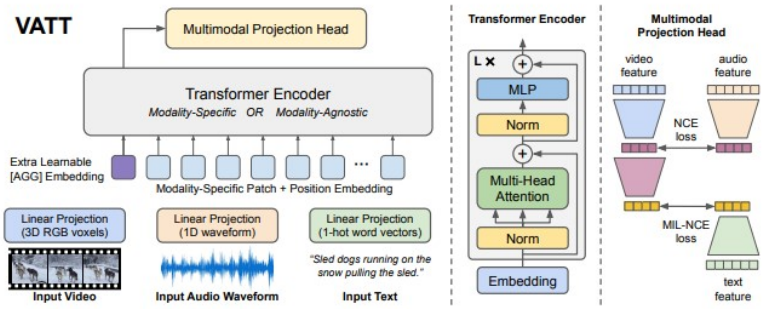


Figure 2.4: The architecture of VATT (left) and the self-supervised, multimodal learning strategy (right), Figure from [1].

2.3.5.3 SSE-FT

The Self Supervised Embedding Fusion Transformer (SSE-FT) is the first transformer that is built from Self-Supervised Learning (SSL) embeddings to represent multiple modalities [97]. Fusing SSL embeddings can be challenging because they come with high dimensionality and long embedding lengths. Between modalities, the embeddings can mismatch in size and sequence length.

To tackle these challenges, Siriwardhana et al. use an attention-based fusion mechanism that can be seen in Figure 2.5. As a first step, features are extracted from the raw data using three pre-trained SSL models, namely RoBERTa and Wav2Vec for text and audio, respectively, and Fab-net for video frames. To effectively represent the modalities, a CLS token, functions as a compressed representation of the embedding sequences. The Inter Modality Attention (IMA) transformer blocks embed one modality's representation with information from other modalities. This layer captures cross-modal information. It operates similarly to self-attention, using the CLS token of one modality as the Query (Q) vector and the embedding sequence of another modality as the Key (K) and Value (V) vectors. There are six IMA transformer blocks, one for each modality pair. In the next step, the embeddings are grouped based on their target modality. After that, to extract the most important information from the target modality, the Hadamard product is taken between CLS tokens of the same target modality. At last, the three representations are concatenated and sent through a fully connected layer to which the softmax function is applied to perform prediction.

2.3.5.3.1 Ablation studies on SSE-FT Siriwardhana et al. conducted an ablation study to better understand the contribution of the different components in SSE-FT[97]. Their study, involved the following ablations on the CMU-MOSEI dataset: Ablation

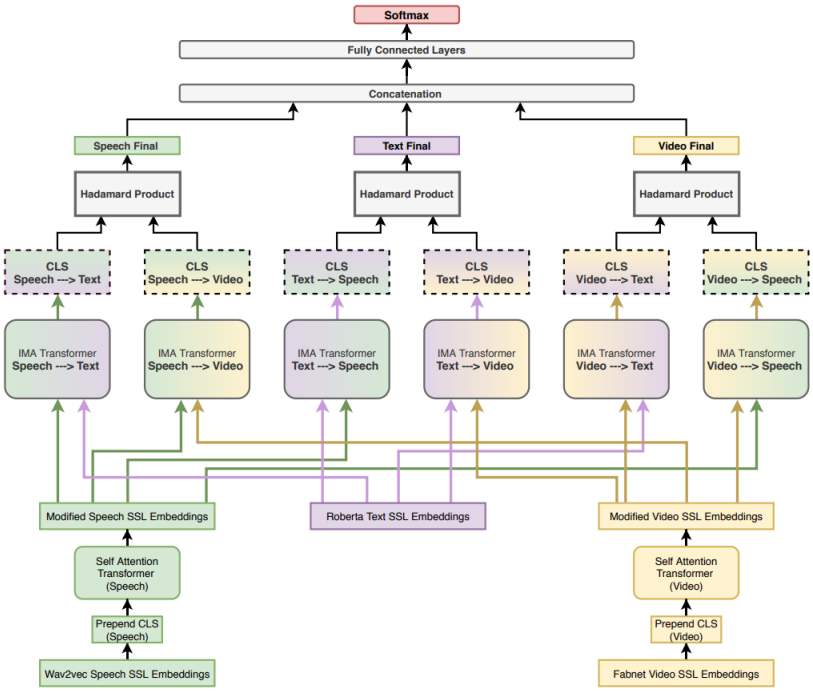


Figure 2.5: The architecture of the Self Supervised Embedding Fusion Transformer (SSE-FT) [58].

study on speech, text, and video input modalities; Examination of speech, text, and video modalities; Analysis of IMA layers (Pre-IMA block); Evaluation of the Hadamard product (Post-IMA block).

For the ablation study on uni-modal input, the authors trained SSE-FT for each individual modality. The CLS token of the modality is extracted after the self attention transformer, and is used to represent the data. This resulted for seven-class accuracy in: text: 47.7%, speech: 43.8%, and video: 43.6%. Moreover, the authors investigated the model's performance using combinations of two input modalities. Here, CLS tokens were taken after the IMA transformer blocks as the final representations. Text and speech inputs demonstrated the highest results for seven-class accuracy: 54.1%. Speech and video inputs yielded the lowest performance metrics: 44.18%. The complete results of their ablation study are shown in Figure 2.6. In the figure, L, A, and V stand for Language, Audio, and Visual. The notation '(h)' indicates that higher values are preferable, while '(l)' indicates that lower values are preferable.

Metric	Acc (7 classes - h)	Acc (2 classes - h)	f1 score (h)	MAE (l)	Corr (h)
Uni-modal Transformers					
Text only	47.7	80.2	80.4	0.636	0.677
Video only	43.6	66.3	65.9	0.729	0.345
Speech only	43.8	67.5	67.8	0.709	0.374
Dual-modal Transformers					
Text and Video	53.9	85.9	85.7	0.543	0.752
Text and Speech	54.1	86	85.8	0.534	0.776
Video and Speech	44.18	68.2	67.8	0.702	0.381
Pre-IMA Fusion Mechanisms					
A+V+T **Three CLS** token concatenation	47.5	81.9	81.8	0.618	0.685
Post--IMA Fusion Mechanisms					
A+V+T **Six CLS** token concatenation	53.3	84.6	84.1	0.567	0.737
Our Final Model (SSE-FT)					
Our Final Model (SSE-FT)	**55.5**	**87.3**	**87**	**0.529**	**0.792**

Figure 2.6: Evaluation results of the ablation studies performed by the authors of SSEFT, Figure from [97].

2.4 Datasets for multimodal emotion recognition

Most of the multimodal models for recognizing emotions are supervised, which means that they require large manually annotated datasets. The annotation process for emotion recognition is a difficult task, since emotions are nuanced and subjective to the annotator. This section further explains the nuances of annotating (multimodal) emotions. Furthermore, multimodal datasets for emotion recognition are highlighted. In table 2.1, an overview of the datasets can be seen, along with their total length, their present modalities, the number of emotion labels, and their year of publication. Datasets that did not include all modalities used in this research, i.e., text, audio, and video, have been excluded from the overview.

2.4.1 Annotating emotions

When annotating a video, an emotion is assigned to each utterance. An utterance can be defined as a complete unit of speech in spoken language. Often, an utterance has the length of a sentence, or the length of the words between speaking pauses. A multimodal utterance includes, for example, the text, audio and video corresponding to the timestamps from this utterance.

The process of annotating emotions to a multimodal utterance is non-trivial due to the subjectivity of annotators, the ambiguity of emotions, and the lack of consistent annotating rules [43]. When analyzing an utterance, there are different emotions that can be observed and labeled. There is not only the speaker's subjective emotion but also the speaker's appraised emotion of a third subject, or even the emotion the annotator feels from observing the speaker. A person can talk about their past or future emotions,

while currently experiencing another emotion. Because of these nuances, in some cases, annotators will not reach agreement as multiple emotions can be assigned. Since often there is no clear distinction and emotions can overlap [43].

The reliability of annotations is typically measured by calculating inter-annotator agreement (IAA), for instance, using the Fleiss' kappa score. A high Fleiss'kappa score means high agreement. Two well-known multimodal emotion datasets achieve fair to moderate agreement: The MELD dataset achieves an overall Fleiss' kappa score of 0.43, while the kappa score for the IEMOCAP annotation process is 0.4 [81]. In most cases, disagreement is resolved by aggregation methods, such as majority voting in MELD [81].

It could be argued that annotating the emotion from a multimodal utterance is somehow more natural than extracting the emotion, e.g., from solely text. This is because theories on emotional categories are based on natural human interaction, as will be described in Section 2.4.1. However, in their study on IAA for emotion recognition in uni-modal versus multimodal utterances, Du et al. found the highest IAA when annotating the text modality alone. Despite each modality offering unique emotion cues, the agreement at the multimodal level was lower compared to text and audio alone. The study also revealed a significant inconsistency in emotion labels across multimodal and uni-modal setups, with nearly half of the instances showing differing emotion labels [36]. This suggests that each modality contributes differently to the perception and labeling of emotions, and learning these inconsistencies could benefit the differentiation of nuanced emotions.

2.4.1.1 Categories of emotion

There are various theories on how to categorize emotions. Basic emotion theory holds that there are basic emotions in humans. These emotions are happiness, sadness, fear, anger, disgust, and surprise [37]. Following this theory, all humans express their emotions from instinct in the same situations in a similar way, producing comparable physiological signals. Other emotions, such as satisfaction or confusion, are believed to be composed of these 6 basic emotions.

On the other hand, the dimensional emotion theory asserts that emotions aren't distinct categories but rather exist on a spectrum, described by varying levels of intensity along different emotional dimensions [85]. These emotional dimensions are primarily valence (ranging from positive to negative) and arousal (ranging from calm to excited states). While multiple approaches to dimensional emotion theory exist, the most commonly adopted model is Russell's 2D emotion model [85]. This model, as envisioned by Russell, visualizes emotions across the valence-arousal spectrum, as depicted in Figure 2.7.

Most multimodal datasets for emotion recognition include both categorical and di-

mensional labels. Combining both emotion label frameworks could offer complementary insights into how emotions are expressed in real life. As categorical labels do identify distinct emotions, they do not capture the varying intensity levels of dimensional labels [14].

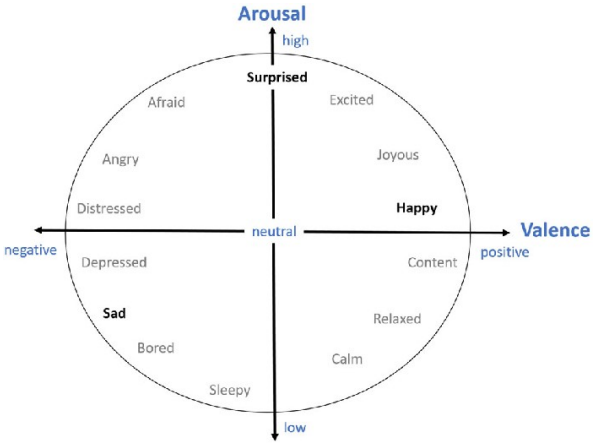


Figure 2.7: The 2D valence-arousal model of emotion proposed by Russell, Figure from [106].

2.4.2 CMU-MOSEI

CMU-MOSEI, the 'Multimodal Opinion Sentiment and Emotion Intensity' dataset, was developed by the Multicomp Lab within Carnegie Mellon University in 2017 [116]. The dataset has 22,777 movie review videos sourced from YouTube, accompanied by 22,856 annotated utterances. The videos are randomly chosen from various topics (250 in total). The dataset is gender-balanced and has a diverse range of speakers (1000 unique speakers). Multiple speakers help a model generalize well over emotional patterns by not solely focusing on individual identities. The annotations are obtained using 3 annotators. Six emotion labels are included for each utterance; angry, happy, sad, surprise, fear, and disgust. Moreover, each utterance is annotated with the scale of emotion within the range of -3 to +3. From very negative -3 to very positive +3. Moreover, the dataset incorporates valence (negative to positive), arousal (passive to active), and dominance (submissive to dominant) scores. The dataset contains the raw data as well as features extracted for all modalities. BERT embeddings are extracted for the text modality, COVAREP and OpenSMILE features are present for the audio modality, and FACET for visual features. While CMU-MOSEI stands out as the prevailing benchmark dataset for MMER, the dataset has some limitations, such as only containing monologues and not having information on multi-speaker interactions.

2.4.3 IEMOCAP

IEMOCAP, the 'Interactive Emotional Dyadic Motion Capture Database', curated by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California, was created out of the need for multimodal datasets capturing human interactions [14]. IEMOCAP is built out of five acting sessions recited from scripts and improvised, each with two actors (one male and one female), mimicking real life interactions. The dataset captures the context of these interactions, which is important to how we express and perceive emotions in everyday life. IEMOCAP spans 12 hours of data collection across ten actors, with 6 emotions such as anger, disgust, fear, sadness, happiness, and, additionally, neutral. In total, the dataset has over 10.000 raw video samples. Valence, arousal, and dominance scores are also included.

2.4.4 MELD

The MELD dataset, 'Multimodal Emotion Lines Dataset', is the first multimodal dataset for emotion recognition that includes 'multi-party' conversations involving text, audio, and video. Meld extents on the Emotion Lines dataset, with textual dialogues from the TV series 'Friends', as the dataset also incorporates video and audio [81]. The dataset is composed of almost 1400 videos, spanning 50 hours in total. The MELD dataset, consisting of 13000 utterances, is split into train, validation, and test folds, 9989 utterances for training, 1109 for validation, and 2610 for testing. Each utterance has an average time span of 3.59 seconds. Furthermore, the dataset includes feature vectors for text and audio. GloVe [77] is used for textual embeddings, and audio features are extracted from the OpenSMILE toolbox [39]. The utterances are annotated with the emotions anger, disgust, sadness, joy, surprise, fear, and neutral. From the Emotion Lines dataset, the annotations were reevaluated, as annotators were to watch the corresponding video to the text. Including multiple modalities while annotating, the MELD annotation process recorded 89 disagreements, significantly fewer than the 2,772 disagreements noted in EmotionLines, which reflects an improved annotation quality achieved with a multimodal dataset.

Despite the significance of the dataset, it has limitations. 'Friends' is a comic TV show and emotions are exaggerated. Moreover, the TV show has various but limited speaking subjects. As can be seen in Figure 2.8, there is an imbalance in the dataset, the 'neutral' label is overrepresented. This can cause problems as it can influence model training towards favoring the neutral class, potentially compromising the model's ability to accurately detect and differentiate other emotions [21].

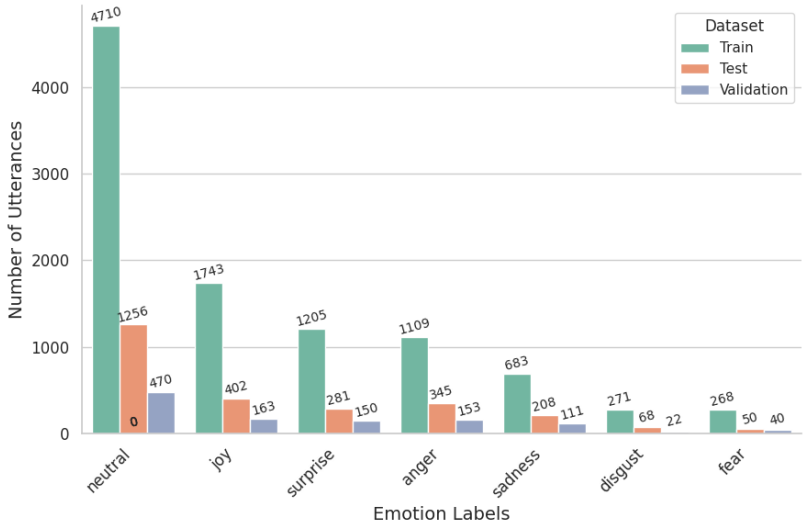


Figure 2.8: MELD: Emotion label distribution across train, test, and validation datasplits.

2.4.5 MEmoR

The dataset MEmoR has videos extracted from the TV show 'The Big Bang Theory' [95]. It has 5,502 videos and 8,536 labeled utterances. In contrast to the datasets mentioned before, MEmoR provides annotations for each speaker and non-speaker. The utterances are labeled using Plutchik's wheel of emotions [79] with 8 primary and 24 more nuanced emotions. The primary emotion labels are joy, anger, disgust, sadness, surprise, fear, anticipation, and trust. OpenSmile is used to extract audio features. A textual representation is extracted with the use of BERT. Facial features and object recognition features are extracted as well. A pre-trained CNN is used for face extraction, and a Facenet model pre-trained on VGGFace2 is used for face recognition [19]. The dataset is finetuned for the seven main characters from the TV series. Moreover, another Facenet model is pre-trained for facial expression recognition.

2.4.6 OMG

The OMG-Emotion, 'One-Minute Gradual-Emotional Behavior' dataset has a total of 420 YouTube videos with a total time of around 10 hours [11]. The videos were automatically selected based on specific search terms related to the term 'monologue'. Each annotator considered the whole video when labeling an utterance, which provides context to the dataset. Arousal and valence are annotated, along with six categorical emotion labels: anger, disgust, fear, happiness, sadness, surprise, and the neutral label.

2.4.7 SEMAINE

The SEMAINE dataset, which stands for 'Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression', was made in 2012 to support research in human-computer interaction [70]. It includes recordings of human subjects in various emotional states, in several modalities, such as text, audio, and video. It also includes physiological signals, which can provide additional insights into emotional states. In the scenario's from the recordings, Sensitive Artificial Listeners (SAL) respond as emotionally stereotyped 'characters'. The dataset includes 80 videos with a total time length of 6 hours. Moreover, it provides 7 categorical emotion labels and values for valence, arousal, intensity, and anticipation.

Dataset	**Length**	**#Videos**	**#Utterances**	**#Modalities**	**#Labels**	**Year**
CMU-MOSEI	1000 hours	22,777	22,856	Text, Audio, Video	6	2017
IEMOCAP	12 hours	302	10,000	Text, Audio, Video	5	2008
MELD	50 hours	1,394	13000	Text, Audio, Video	7	2018
MEmoR	32 hours	4,500	8,536	Text, Audio, Video	9	2020
OMG	10 hours	420	2400	Text, Audio, Video	7	2018
SEMAINE	6 hours	80	-	Text, Audio, Video	7	2012

Table 2.1: Overview of datasets for MMER including the video, audio, and text modalities.

2.5 Interpretability in multi-modal models

Multimodal models have been using the combination of multiple modalities to enrich the model's pattern recognition abilities. Using advanced fusion techniques, these models, improve on their uni-modal predecessors. However, neural nets have complex hidden layers from which we have little or no understanding. This makes the decision-making process and internal states of neural networks 'black box'.

From other simpler machine learning techniques, such as linear regression and decision trees, we know how the decision-making process works. The coefficient of a feature in the equation of linear regression is a clear representation of the importance of this feature. Decision trees are structured to first split on the most informative feature, analyzing the splits of the tree, we can understand how the algorithm came to its decision [89].

The field of XAI , which is a term invented by the Defense Advanced Research Project Agency (DARPA), aims to improve our understanding of more black-box models such as neural networks [49]. These non-linear models are not inherently interpretable and, therefore, ask for a more complex 'post-hoc' approach. Interpretable representations of the model's decision making process on the input can be made. In the context of

text classification, for instance, interpretability can be achieved with binary vectors representing the presence or absence of specific words, even though the underlying classifier may work with more complex features like word embeddings. Multiple reliable interpretability methods that can produce understandable representations have been proposed and will be discussed in Section 2.5.1.

2.5.1 Interpretability methods

In this section, four prominent interpretability methods are explored: SHAP (SHapley Additive exPlanations) in Section 2.5.1.1, LIME (Local Interpretable Model-agnostic Explanations) in Section 2.5.1.2, the attention mechanism in Section 2.5.1.3 and prototype-based methods in Section 2.5.1.4. These methods offer diverse approaches to uncovering the inner workings of multimodal models.

2.5.1.1 SHAP

The SHAP (SHapley Additive exPlanations) method draws its inspiration from the concept of Shapley values, which originate in the fields of economics and game theory [94]. The goal of this method is to fairly distribute rewards from a set of games to all the players. In the context of machine learning, the SHAP method achieves this by linking the model's features with the 'players' who are set to receive the rewards [65]. Thus, when making a prediction, the SHAP method breaks down the prediction into components, focusing on the contributions of each feature. This process allows for the calculation of the individual contributions of each feature to the prediction [57].

Shapley values have four defining properties required for a fair payout, which all add to the interpretability of a model: "1) Efficiency: the contributions of all players sum up to the model outcome; 2) Symmetry: any two players that contribute equally are assigned the same payout; 3) Dummy: a non-contributing part is assigned zero value; and 4) Additivity, enabling us to simply average the Shapley Values to determine the total player contributions in a game with combined payouts (e.g., the two halves of a soccer match, or ensembling of decision trees)" [65].

SHAP is model-agnostic, it means, in the context of machine learning, that the technique can be applied to various models, as it does not use the properties of the model architecture to generate explanations. Therefore, this technique is widely used for interpreting black-box models.

Apart from SHAP having a grounded theoretic foundation as an interpretability method, the SHAP library [1] also provides a wide range of visualization plots to support its explainable power. Wang et al. use Shapley values to visualize the contribution of features in the metadata (e.g., age, gender, etc.) used in an interpretability-based

[1] https://shap.readthedocs.io/en/latest/

multimodal CNN for skin lesion diagnosis [109]. Zhang et al. use SHAP visualizations to detect referable diabetic retinopathy, showing the use of explainable deep learning for predictive healthcare tasks [118]. A complete summary of the SHAP visualization plots will be given in Section 2.5.1.1.1. Parcalabescu et al. use SHAP to analyze pre-trained vision-language encoders on their modality contribution. The authors find that different models have different dominating modalities on the same task with the same dataset [75]. Their approach will be elaborated in Section 2.5.1.1.2.

2.5.1.1.1 SHAP visualization plots Visualizations of Shapley values can be effective explainatory tools for understanding the patterns and dependencies in a model's prediction process. A short summary and application example of each visualization plot currently provided by the SHAP library will be given. The visualizations can be seen in Appendix A.

2.5.1.1.1.1 Bar plot The bar plot visualizes the average Shapley values for each feature. These features can be analyzed globally (over all samples) as well as locally (for one sample). The bar plot also enables the visualization of cluster importance, in cases where features are redundant with each other (redudant meaning that a model could use either feature and still get the same accuracy). An example of the SHAP barplot can be seen in Figure A.11.

2.5.1.1.1.2 Beeswarm plot The beeswarm plot shows the distribution of the data samples for the top contributing feature. The plot can help gain insight into how feature importance varies across different data points. The plot provides a summary of the overall effect of each feature on the model's output. An example of the SHAP beeswarm plot can be seen in Figure A.12.

2.5.1.1.1.3 Violin plot The violin plot, like the beeswarm plot, shows the distribution of the datapoints for all features, providing a summary. It, however, adds the central tendency, spread, and symmetry of the data distribution. An example of the SHAP violin plot can be seen in Figure A.23.

2.5.1.1.1.4 Decision plot The decision plot is effective for showing how more complex models arrive at their predictions (which decisions are made). It can be used to show the cumulative effect of the features. For binary classification, it shows how the output changes as each feature varies. For multi class classification, the plot shows how the model's decision varies across different classes. The decision plot can also be used to compare the decision behavior of different models. Moreover, it can be used to group

observations with similar prediction paths to detect outliers. An example of the SHAP decision plot can be seen in Figure A.24.

2.5.1.1.1.5 Heatmap plot The heatmap plot shows all samples on the x-axis and all features below on the y-axis. By visualizing each feature value for each sample, the heatmap plot shows a full picture of the variations in different features across the dataset. An example of the SHAP heatmap plot can be seen in Figure A.25.

2.5.1.1.1.6 Dependence Scatter plot The dependence scatter plot can show the models learned dependencies for the data. In the scatter plot, the range of feature values can be plotted against the x-axis and their corresponding Shapley values along the y-axis. This way, the scatter plot can be used to analyze the interaction between the feature and the prediction as the feature value changes. More features can be added to examine their dependencies. An example of the SHAP scatterplot can be seen in Figure A.26.

2.5.1.1.1.7 Force plot The force plot shows the feature importance for a single sample as a series of bars. Each bar represents a feature, and its length and direction indicate the influence (positive or negative) on the prediction. An example of the SHAP force plot can be seen in Figure A.37.

2.5.1.1.1.8 Waterfall plot The waterfall plots, like the force plot, show the importance of a feature for a single sample. However, features are shown along the y axis, and the cumulative contributions of the features are shown on the x axis. An example of the SHAP waterfall plot can be seen in Figure A.39.

2.5.1.1.1.9 Text plot The text plot is used to specifically interpret the contributions of individual words to the prediction of a text-based machine learning model. It includes a force plot, which shows the words instead of features. In addition, it shows the text input and marks the words with red and blue, indicating their positive or negative contributions to the prediction. An example of the SHAP text plot can be seen in Figure A.38.

2.5.1.1.1.10 Image plot The image plot is perhaps the best known plot for visualizing Shapley values. It visualizes how different parts of an image contribute to the model's decision-making process. It can be used besides object recognition to analyze the influence of the presence and position of objects to the model's output. An example of the SHAP image plot can be seen in Figure A.310.

2.5.1.1.2 Multi-Modal SHAP Parcalabescu et al. extended SHAP to involve more modalities, i.e., image and text, through the introduction of MM-SHAP (Multi-Modal SHAP). The authors introduce MM-SHAP as a metric to measure the multimodal degree, which is the degree to which modalities are used in model predictions. Measuring the multimodal degree can test for uni-modal collapse, whether a multimodal model has ineffective fusion and relies on only a single modality. Like uni-modal SHAP, MM-SHAP is performance agnostic, meaning that its outcomes are independent of the model's performance. This characteristic is preferred because, unlike performance-based methods that overlook false predictions, MM-SHAP considers all predictions. Even false predictions provide valuable insights into how the model processes features, making them useful for analyzing feature contributions. Moreover, using a performance-agnostic metric allows for the measurement of the multimodal degree in situations where model accuracy is low.

The multimodal degree as proposed by Parcalabescu et al. is computed as follows: Shapley values are computed for the multimodal transformer model at the prediction time for all the samples in the test set. The complete input to the model is represented by tokens. As the model has multimodal input, each modality has a unique token representation. MM-SHAP uses so called superpixels to represent the visual modality, and for the textual modality, each token represents a word, as shown in Figure 2.9.

The input to the models consists of p tokens, including text and image tokens. First, subsets $S \subseteq \{1, \ldots, n\}$ of tokens are created, forming a group towards the model prediction Val(S). The Shapley value (ϕ_j) for a token j is calculated using the formula:

$$\phi_j = \frac{1}{\gamma} \sum_{S \subseteq \{1,\ldots,n\} \setminus \{j\}} \frac{[val(S \cup \{j\}) - val(S)]}{\binom{n-1}{|S|}} \tag{2.1}$$

where $\gamma = \binom{|S|}{n-1|S|} \cdot \frac{p!}{n!}$ is the normalizing factor accounting for all possible combinations of choosing a subset S. The number of potential coalitions grows exponentially with the number of tokens masked, denoted as p, resulting in ($\gamma = 2^{2p+1}$). Due to the impracticality of computing Shapley values for all possible subsets, the 'Monte Carlo' approximation is employed [65]. This involves randomly sub-sampling $n = 2p + 1$ subsets to estimate the Shapley values.

Now the textual contribution is defined as Φ_T, the image contribution as Φ_I towards a prediction as the sum of (absolute) Shapley values of all textual and visual tokens for a pre-trained multimodal transformer with n_T text tokens and n_I image tokens:

$$\Phi_T = \sum_{j=1}^{n_T} |\phi_j| \tag{2.2}$$

$$\Phi_I = \sum_{j=1}^{n_I} |\phi_j| \tag{2.3}$$

The absolute value is considered, and not the sign (positive or negative) of the influence of a token, as the focus is on measuring whether a token is influential in a modality regardless of the direction it pushes the prediction. MM-SHAP can also be defined as a 'proportion of modality contributions', which allows for assessing a model's textual degree (T-SHAP) and visual degree (V-SHAP).

$$T\text{-SHAP} = \frac{\Phi_T}{\Phi_T + \Phi_I} \tag{2.4}$$

$$V\text{-SHAP} = \frac{\Phi_I}{\Phi_T + \Phi_I} \tag{2.5}$$

Parcalabsecu et al. test their multimodal framework for interpretability on four tasks: Image-Sentence Alignment (ISA), Visual Question Answering (VQA), Visual Question Answering with Balanced Datasets (GQA) and Vision and Language for Scene Understanding (VALSE). Figure 2.9 illustrates the proposed framework on the task of ISA and shows the contribution of the tokens to the ISA scores, together with the textual-contribution. The contribution of the visual modality is computed as 100 - T-SHAP.

The scalability of the method proposed by Parcalabescu et al. is a notable concern. The computational complexity of the method can increase significantly as more modalities are added and more detailed interpretability is desired. The superpixels used in MM-SHAP to represent the visual modality are limited in their explanatory power, while they can overlap multiple objects, and pixels residing within neighboring patches often share semantic relevance. Tasks such as image-sentence alignment and visual question answering rely on the understanding of concepts such as objects and their context.

To address this limitation, more fine-grained tokens can be created to better represent individual objects and their contexts. However, this approach increases computational complexity significantly. Cafagna et al. propose a solution by using the visual backbone of a transformer model to generate tokens for images that represent meaningful areas, such as individual objects [16]. This approach aims to reduce the number of visual input features as well as construct more semantically valid explanations. Although this approach yields more semantically relevant explanations, it violates the feature independence assumption of Shapley Values, assuming that the features are disjoint and do not correlate. Applying this approach for calculating modality contribution could give inconsistent results as the regions generated with Deep Feature Factorization (DFF) can overlap. However, Cafagna et al. show that the overlap does

not significantly affect the final segment contribution for a sample. When aggregating over all samples in a dataset to calculate modality contribution, the use of DFF could potentially lead to an overestimate of the contribution of the visual modality. In contrast, using superpixels, which do not overlap, would avoid the risk of double counting a region's contribution.

Other approaches addressing the computation of Shapley coefficients in image data are h-shap, which explores a hierarchical partition of the input image [102]. This method begins by broadly partitioning the image and then focuses on further subdividing only the areas deemed important.

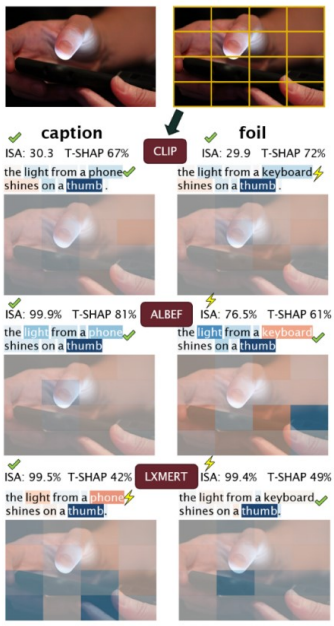


Figure 2.9: MM-SHAP: This figure illustrates the ISA score for six different VL models with their respective T-SHAP values, represented as percentages. Blue tokens contribute positively to a high ISA, while red tokens lower the ISA. Correct and incorrect alignments are marked, with correct alignments highlighting tokens contributing positively to aligning the image and caption, and incorrect alignments indicating a negative contribution [75]

.

2.5.1.2 LIME

The 'Local Interpretable Model-agnostic Explanations' or LIME is an algorithm for providing insights into the predictions made by any classifier or regressor [84]. LIME, just like SHAP, is model-agnostic, meaning that the mechanism treats the original model as a black box. LIME works as follows: for each sample, LIME generates altered

versions of the sample by making small, controlled changes. The idea is to explore how the model behaves when the input data changes slightly. For example, in Figure 2.10, at every altered version, some connected pixels are left out, resulting in a changed classification of the sample. If the absence of certain connected pixels consistently leads to a change in the model's classification for the altered versions of a sample, it suggests that these pixels are important for the model's decision-making in a local context. LIME is more used for local features but could also be used for global feature interpretation and assessing model behavior.

DIME, an approach based on LIME, aims to disentangle explanations from LIME and can analyze the impact of each modality independently [66]. In this approach, LIME is executed for one modality at a time, while keeping the inputs to all other modalities constant and only perturbing the inputs to the selected modality. This process allows the authors to analyze the impact of each modality on the explanation generated.

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 2.10: LIME: Explaining the predictions for the top 3 predicted classes (b, c and d) for the original image in a for image classification [84]

.

2.5.1.3 The Attention Mechanism

Another method for interpreting multimodal predictions is leveraging the attention mechanism in transformers. As previously explained in Section 2.3.4, the attention layer determines what part of the input the focus should be on. The attention values, therefore, tell us what features are important to the prediction: high attention values correspond with high feature importance. It is noteworthy, that attention values allow for positive-only relevance assessments (while Shapley values can also make negative-relevance assessments) [25].

Transformers can have multiple layers of self and co-attention. For each of these attention layers, attention maps can be visualized to provide valuable insight into the inner workings of a model. Chefer et al. propagate through attention layers to produce relevancy maps for each of the interactions between the input modalities in the network [25]. They test their approach for models using self-attention and co-attention. An

example of the relevance maps produced with their method can be seen in Figure 2.11. Relevance for images is given by multiplying each region by the relative relevancy [48]. In contrast to this technique, raw attention maps regard only the last layer's attention map for feature importance.

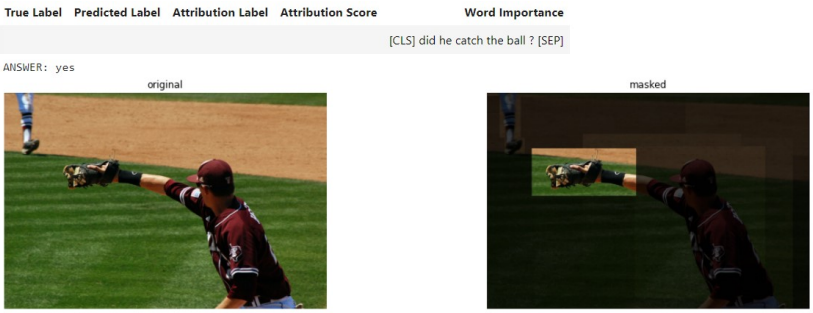


Figure 2.11: An example of a relevance map showing the focus of the model for VQA, Figure from [25].

Another technique for visualizing attention values is 'Gradient-weighted Class Activation Mapping', also called Grad CAM's attention maps [93]. An example of a Grad CAM heatmap is shown in Figure 2.12. Important regions have higher attention scores and are therefore displayed as red in the heat map. The above-mentioned methods are

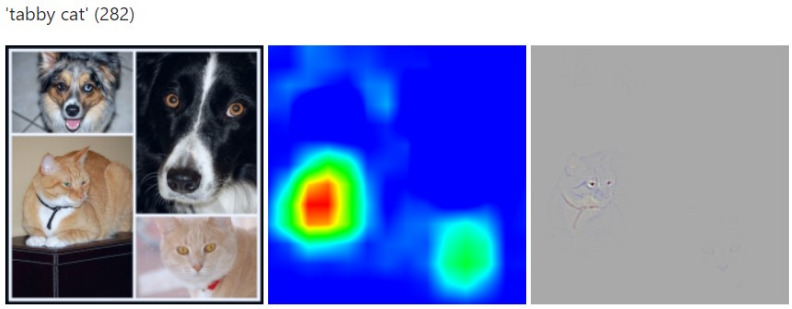


Figure 2.12: An example of a Grad CAM heatmap showing the focus of the model on the tabby-cat, Figure from [93].

a form of post-hoc attention, meaning that the attention values are extracted after the training phase, and so, parameters are already set. Contrary to post-hoc attention, trainable attention refers to the method where attention weights are learned during training [48].

Attention scores can also be used to visualize and calculate modality contributions. Showing how modalities contribute to a prediction, given a certain task and dataset, can provide insights into how the model integrates the input from different sources.

For instance, Cao et al., in their research, have utilized attention weights from pre-trained visual-language models to visualize the significance of different modalities [18]. Additionally, attention schemes have been employed to assess the contribution of each modality to the outcome in emotion recognition [63]. The authors make use of the [CLS] and [SEP] tokens present in the transformer architecture. Information from all modalities is represented in the [CLS] token through self-attention; thus, the degree of attention of the [CLS] token over each modality can be calculated to investigate the contribution of the modalities.

Furthermore, apart from visualizing feature relevance in self-attention and co-attention layers, cross-modality correlation can be analyzed by computing the attention scores between two modalities by cosine similarities [105]. Tsai et al. propose a method that fuses the values from both modalities based on the attention scores. The authors show both global and local interpretability in their model. For each test sample, they show the contribution from each single modality and all of their combinations.

Although attention provides interpretability in a model's decision-making process, there is no consensus yet on whether attention yields explanatory power. For example, researchers investigated and found that the perceptions of a model and a human do not focus on the same areas while providing the same output [47]. However, the methods used in their study are criticized, and it was found that it is not excluded that attention can account for interpretability [112]. Even if the model and human perception don't align perfectly, understanding what the model focuses on can provide insights into the decision-making process of these black-box models.

2.5.1.4 Prototype-based interpretability methods

Prototype-based interpretability methods rely on a set of typical representatives known as prototypes. This approach resembles clustering algorithms like K-Means, in which the centroids of clusters can serve as prototypes. In their work, Zinemanas et al., propose a prototype-based interpretable model for audio classifications [122]. Their model has a prototype layer that stores several prototypes, which are representatives of each class. The layer outputs a similarity measure for each input sample to each prototype. The similarity measures of the input samples to the prototypes can be analyzed to make intuitive explanations.

This method stands out because it provides a clear window into the model's decision-making process, and eliminates the necessity for an additional interpreter, as required, for instance, with attention maps. Chen et al. propose a prototypical part network (ProtoPNet), which can be seen in Figure 2.13. Their method dissects an image and finds prototypical parts to combine evidence from the prototypes to make a final classification [26]. Kim et al. present an interpretable vision transformer neural tree (ViT-NeT). Their method uses a ViT to achieve high-level classification performance and a

neural tree that acts as a discriminant decoder, interpreting the decisions of the ViT and then routing the images hierarchically [53]. This opens the decision-making process of the model. Zhang et al. propose a method that selects sub-sequences that represent concepts of the input sequence as prototypical parts [120].

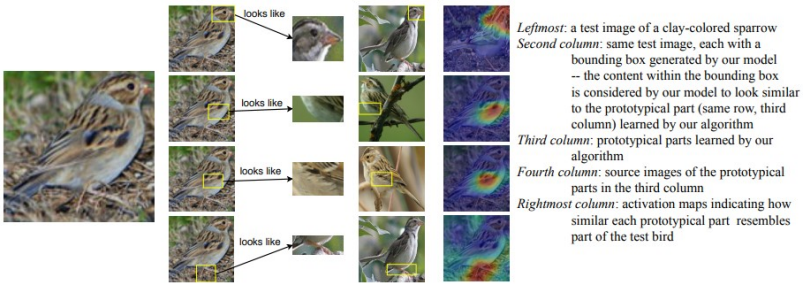


Figure 2.13: An example of the classification of a bird by ProtoPnet. The image is divided into parts, which are each linked to learned prototype parts belonging to a source image. The rightmost column shows the activation maps, indicating the similarity to the prototype. Figure from [26].

Chapter 3

Methodology

This chapter presents the methodology and details of the experimental setup of the current research. An overview of the research is given in Section 3.1. The dataset used for finetuning and evaluating the multimodal transformer model will be explained in Section 3.2. The implementation of the multimodal transformer SSE-FT will be explained in Section 3.3. Moreover, the modification and implementation of the interpretability method MM-SHAP will be explained in detail in Section 3.4. In Section 3.6, the implementation and evaluation of the multimodal model on the Sound & Vision archive will be elaborated. Additionally, designs for visualizing the results from the interpretability are presented in Section 3.5.

3.1 Research overview

For a more structured overview of the current research, this section provides the research steps and refers to the sections that discuss them accordingly.

1. The MELD dataset undergoes pre-processing to meet the specific requirements of the model SSE-FT. An outline of the MELD dataset is given in Section 3.2.
2. SSE-FT is fine-tuned and evaluated on the MELD dataset as described in Section 3.3.
3. The MM-SHAP framework for multimodal interpretability is implemented for SSE-FT. The process of modifying MM-SHAP is described in Section 3.4.
4. The modified MM-SHAP is used to calculate modality contributions for the MELD dataset, for each emotion label, and on a sample level, as described in Section 3.4.
5. Visualizations are designed to clarify the Shapley values at the sample level. The method for creating these visualizations is specified in Section 3.5.

6. SSE-FT and MM-SHAP are implemented in the digital pipeline of the Institute of Sound & Vision as described in Section 3.6.
7. The modified MM-SHAP for emotion recognition is implemented and evaluated on a corpus from the Sound & Vision archive as described in Section 3.6.
8. The outcomes of the modified MM-SHAP as well as the visualizations of the outcomes are given in Chapter 4 and analyzed in Chapter 5.

3.2 The MELD dataset

As previously outlined in Section 2.4, multiple multimodal datasets are proposed for emotion recognition. Each has different features and characteristics, as they are made for different applications within the field. In the current research, the dataset MELD was selected for the following reasons: The Sound & Vision archive contains interviews, talk shows, and actuality programs featuring multiple individuals in a single frame. MELD stands out as the only dataset specifically designed to handle such multi-party interactions. Moreover, MELD has raw data available as well as pre-extracted features for all modalities: text, audio, and video. For the S & V archive, the model is applied to raw videos, and therefore the model that is implemented needs to be able to handle these raw videos as input. MELD showed compatibility not only with the transformer model SSEFT, but also with the other models (VATT and MULT) considered during the model selection process, as these transformer models require raw input data, and notably, SSE-FT has been evaluated on the MELD dataset. Moreover, the modalities in the dataset are already pre-aligned with time stamps, so the model does not have to perform this alignment. The MELD dataset is split into train, validation, and test folds[1]. The train fold consisting of 9989 utterances is used for the finetuning of SSE-FT. The test fold of 2610 samples is used for evaluation. A more detailed description of the properties of the MELD can be read in Section 2.4.4.

3.3 Implementing SSE-FT

This section describes the methodology for the implementation of the multimodal model SSE-FT. First, the metrics used to evaluate the performance of SSE-FT are given in Section 3.3.1. In Section 3.3.2, the procedure of finetuning SSE-FT on the MELD dataset is explained. Lastly, the ablation studies done to evaluate the multimodality of SSE-FT are described in Section 3.3.2.1.

[1] https://affective-meld.github.io/

3.3.1 Metrics for evaluating the performance of SSE-FT

For a complete evaluation of the performance of SSE-FT, the following metrics are used: 7-class accuracy, precision, recall, and the F1 score. The F1 score is the harmonic balance of precision and recall [90]. Moreover, dual accuracy, precision, recall, and the F1 score are calculated to evaluate the model's ability to recognize emotion versus non emotion. For clarity, the computation of dual accuracy for distinguishing between a neutral class (N) and all other emotion classes (non-neutral, O) is provided:

1. **Definition of metrics**:
   - True Positives (TP): Samples correctly classified as a particular class.
   - True Negatives (TN): Samples correctly classified as not belonging to a particular class.
   - False Positives (FP): Samples incorrectly classified as belonging to a particular class.
   - False Negatives (FN): Samples incorrectly classified as not belonging to a particular class.
2. **Calculation of both accuracies**:
   - For the neutral class (N):
   $$\text{Accuracy}_N = \frac{TP_N + TN_N}{TP_N + TN_N + FP_N + FN_N}$$
   - For the non-neutral class (O):
   $$\text{Accuracy}_O = \frac{TP_O + TN_O}{TP_O + TN_O + FP_O + FN_O}$$
3. **Dual accuracy calculation**:
   $$\text{Dual Accuracy} = \frac{\text{Accuracy}_N + \text{Accuracy}_O}{2}$$

3.3.2 The finetuning procedure for SSE-FT

In Section 2.3.5, the architecture of SSE-FT, among other considered transformer models has been outlined. Apart from SSE-FT, the multimodal models, VATT described in Section 2.3.5.2 and MulT 2.3.5.1 were considered. However, after implementation efforts, VATT and MulT appeared to have deprecated code bases. Hence, the model SSE-FT was chosen to be implemented and analyzed with an interpretability framework. Even though SSE-FT has been evaluated on the MELD dataset, pre-trained

checkpoints for their best model on the MELD dataset are not included in the code base. Hence, finetuning on the MELD dataset is necessary. The implementation of the model SSE-FT is available on GitHub[2]. As described in Section 2.3.5.3, SSE-FT initially extracts features from pre-trained SSL models, namely RoBERTa, Wav2Vec, and Fabnet. Checkpoints for these pre-trained models are available. The MELD training split, consisting of 9987 utterances, is used for finetuning.

For finetuning, first, text files are tokenized using RoBERTa, which is integrated into the Fairseq library. For the audio WAV files, the waveforms are converted to tensors and saved as .pt files. The pre-trained SSL models, RoBERTa large[3], Wav2Vec2[4], and Fabnet[5], are downloaded. SSE-FT is validated on the test split with 2610 utterances. In their experiments with the MELD dataset, the authors perform a basic grid search for hyper-parameter tuning. These settings are followed in the finetune procedure and can be found in Table 3.1. In the first training effort, fp16 was activated, for a computational speedup. In the second effort, fp16 was turned off to possibly improve performance. The performance results from SSE-FT can be found in Section 4.1.

Hyper-parameter	**Value**
Batch Size	32
Epochs	20
Number of Attention Blocks	1
Number of IMA Blocks	1
Number of Self Attention Heads	2
Number of IMA Heads	2
Dropout Rate	0.1
Initial Learning Rate	3.00E-04
Learning Rate Scheduler	Polynomial Decay
Training Hardware	NVIDIA A100 Tensor Core GPU

Table 3.1: Original hyper-parameters for finetuning SSE-FT on the MELD dataset [97].

3.3.2.1 Ablation study

In the current research, SSE-FT is trained and evaluated on the MELD dataset. An ablation study is conducted on the speech, text, and video input modalities for the MELD dataset, to better understand how well the model can capture information from each modality individually, without being influenced by the presence of other modalities. The model is also evaluated using the dual input of the audio and video modalities,

[2] https://github.com/shamanez/Self-Supervised-Embedding-Fusion-Transformer
[3] https://github.com/facebookresearch/fairseq/blob/main/examples/roberta/README.md
[4] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
[5] https://www.robots.ox.ac.uk/~vgg/research/unsup_learn_watch_faces/fabnet.html

to assess its performance without the SSL embeddings from RoBERTa. Moreover, the results from the ablation study can be compared to the overall modality contribution values obtained from the modified MM-SHAP. The method for the ablation study is as described in Section 2.3.5.3.1. The CLS token of the modality is extracted after the self attention transformer and is used as the final representation. Hence, the model is trained on the individual modalities text, audio and video and the dual modality audio-video. Results from the ablation study can be read in Section 4.1.1.

3.4 Modifying the multimodal interpretability method MM-SHAP

In Section 2.5.1.1, the workings of the model-agnostic interpretability method SHAP have been explained. SHAP has been extended to involve more modalities, i.e., image and text, through the introduction of MM-SHAP, as discussed in Section 2.5.1.1.2. In the current research, this approach is adopted to explore the modality contribution of the chosen multimodal transformer, SSE-FT, in the context of emotion recognition. Notably, in the proposed method, MM-SHAP is tailored to include the audio modality as well, analyzing the text, video, and audio modality together. The current section defines the calculation of Shapley values and the modality contribution, including all three modalities, and the experiments using these calculations (Section 3.4.0.1) and explains the implementation to the model SSE-FT (Section 3.4.1.1).

3.4.0.1 Calculating Shapley values and the multimodal degree

Shapley values are computed for the multimodal transformer models at prediction time on the test dataset. As previously explained in Section 2.5.1.1.2, first subsets $S \subseteq \{1, \ldots, n\}$ of tokens are created to form a group towards the model prediction val(S). And then, the Shapley value (ϕ_j) for a token j is calculated using Equation 2.1. Originally, only the text and image degrees were measured. However, in this research, MM-SHAP is augmented with the audio modality.

To determine the multimodal degrees, the textual contribution is defined as Φ_T, the image contribution as Φ_V, and the audio contribution as Φ_A towards a prediction as the sum of Shapley values of all textual, visual, and audio tokens for a pre-trained multimodal transformer with n_T text tokens, n_V video tokens, and n_A audio tokens. The contribution of the audio modality Φ_A is computed the same as Φ_T and Φ_V:

$$\Phi_A = \sum_{j=1}^{n_A} |\phi_j| \tag{3.1}$$

As previously discussed, MM-SHAP can also be defined as a proportion of modality contributions. Similar to the model's textual degree (T-SHAP) and visual degree (V-SHAP), the audio degree (A-SHAP) can be computed as:

$$A\text{-SHAP} = \frac{\Phi_A}{\Phi_T + \Phi_I + \Phi_A} \tag{3.2}$$

This formulation accounts for contributions for all three modalities.

3.4.1 Experiments with the modified MM-SHAP

From the calculations in Section 3.4.0.1, in theory, multiple experimental questions could be answered. The following questions can provide interpretability to any multimodal model for emotion recognition, and are used to create interpretability and analyze the robustness of the model SSE-FT fine-tuned on the MELD dataset:

1. *What are T-SHAP, V-SHAP, and A-SHAP on dataset level?*
   The contributions of the text, video, and audio modality to the predictions averaged over the MELD test set are calculated to analyze the multimodality of SSE-FT. How completely and effectively does SSE-FT use the information from each part of the input?

2. *How do T-SHAP, V-SHAP, and A-SHAP vary for each emotion label?*
   The contribution of the modalities to emotional classes is analyzed to understand how SSE-FT uses each modality to discern emotions. Are some emotion classes best recognized by a certain modality?

3. *What do samples with either text, video, or audio as the highest contributing modality have in common?*
   Instances where either text, video, or audio is identified as the modality with the highest contribution to the model's predictions are analyzed. The patterns that are shared among these samples can provide insights into which aspects of each modality are most important for recognizing emotions for SSE-FT.

4. *What insights can be derived from T-SHAP, V-SHAP, and A-SHAP when samples are misclassified?*
   To better understand the factors within a multimodal model that lead to incorrect predictions, the Shapley values of misclassified samples are examined. In the case of an incorrect prediction, which label is predicted? Are these emotion classes more often confused? Can a change be noticed in the degree of multimodality? This could suggest that SSE-FT turns to a certain modality when the confidence in a sample prediction is low.

Zooming into each modality, characteristics within the modalities can be analyzed to find out what influence they have on recognizing emotions. To calculate the modality contribution, the input for each modality has been split up into representative tokens. Analyzing the Shapley values computed for these tokens can give information on how a multimodal model and SSE-FT trained on the MELD dataset specifically, recognize emotions. From the token representation further described in 3.4.1.1, the following questions can be analyzed:

1. *What parts of a sentence are important for model prediction?*
2. *What specific time points or intervals of the audio have an impact on the prediction?*
3. *What parts of the video within a frame have an impact on the prediction?*
4. *Is the time dimension of the video input of importance to the model prediction?*

3.4.1.1 MM-SHAP with SSE-FT

As previously discussed, the method of MM-SHAP relies on perturbation, involving the selective masking of input tokens to observe changes in the model's output. The perturbation process uses a masker function, allowing specific tokens from modalities to be shut off. For the method, three things are important: To define how the input for a multimodal model is formatted and so which parts can be masked (Section 3.4.1.1.1), to define for each modality how each input token is represented (Text, Section 3.4.1.1.2, audio Section 3.4.1.1.3, and Video Section 3.4.1.1.4), and to which value the influence of the tokens is calculated (Section 3.4.1.1.5).

3.4.1.1.1 Masker function Firstly, in the masker function, it is defined what the model input looks like in general, even though some tokens are masked. Namely, some tokens should not be masked, as they convey important information. These tokens are the CLS and the separator. Similarly, the padding tokens do not get masked, as including them would not make sense, since these tokens are unimportant to the model outcome. The model SSE-FT has a CLS token for the text input (identified as the first token '0'); it appends CLS tokens for audio and video later in the model. SSE-FT does not use separator tokens. The padding tokens are 0 for audio and text and -1 for video. Hence, the text CLS token and padding tokens get excluded from masking.

3.4.1.1.2 Text modality Having established which parts of the input should not be masked, the next step involves determining how to mask the relevant parts for each modality. Since each modality is represented differently, they require different masking

approaches. The representation of the text input is as follows: each token represents a single textual element in a sentence. Tokens for the textual modality consist of tokens after the tokenization by roBERTa large.

Tokenization with roBERTa includes punctuation as well as the sentence start and end tokens: <s>and </s>. Punctuation and special tokens matter for the context, tone, and nuance of communication. For instance, an exclamation mark might denote excitement or emphasis, while question marks could suggest hesitation or surprise. Therefore, all textual elements have to be included in the analysis of the text modality, as all of them are used by SSE-FT.

An example of the process of masking the text is given in Figure 3.1. The text is first tokenized, and then SHAP masks every combination of tokens to calculate their contribution. For example, consider the tokenized utterance '<s>I feel happy ! </s>', individual tokens are masked '<s>[MASK] feel happy ! </s>', but also pairs '<s>[MASK] [MASK] happy ! </s>', and every other combination '<s>[MASK] [MASK] [MASK] [MASK] </s>'. This process continues for all possible combinations of tokens.

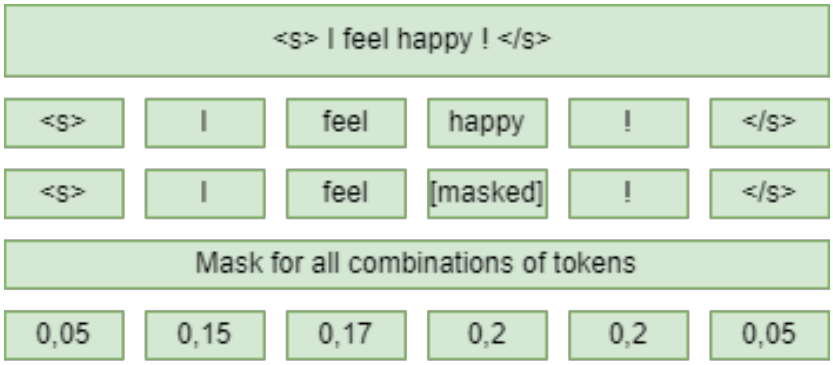


Figure 3.1: An example of the process of masking tokens for the text modality. For clarity, the full words are shown instead of the tokenenized sentence.

3.4.1.1.3 Audio modality For audio, the tokens are half a second long during waveform segments. To construct these audio segments for each audio file, the number of values per token is calculated as the duration of a token (0.5 seconds) divided by the duration of a sample.

$$\text{Values per token} = \frac{0.5 \text{ seconds}}{\text{Duration of a sample}} \tag{3.3}$$

The duration of a sample is calculated as 1 divided by the sample rate.

$$\text{Duration of a sample} = \frac{1}{\text{Sample rate}} \tag{3.4}$$

Combining both equations gives:

$$\text{Values per token} = 0.5\text{Sample rate} \tag{3.5}$$

This results in a vector of audio tokens that represent segments of half a second. As the total number of samples in the audio vector might not be perfectly divisible by the values per token, there is a remainder smaller than half a second. These remaining samples are distributed evenly across all tokens. An example to illustrate this process is given.

- Suppose the sample rate is 16,000 samples per second.
- The values per token would be $0.5 \times 16000 = 8000$ samples.
- If the audio file has a total of 32,050 samples, there would be 4 tokens of 8,000 samples each and a remainder of 50 samples.
- These 50 samples are then distributed evenly across the 4 tokens, adding approximately 12-13 samples to each token.

An example of the masking process of the audio is given in Figure 3.2.

3.4.1.1.4 Video modality The video input is shaped as 1, 3, 300, 256, 256, batch size, channels, number of frames, width, and height, respectively. For each frame, the 256 x 256 patches are divided by 16. As can be seen in Figure 3.4. Each video is represented as 16 tokens, each representing a patch within a 4 x 4 grid. There are 4 rows, and the first row has tokens 1, 2, 3, and 4. The second row has tokens 5, 6, 7 en 8 etc. For each frame in the video, the patches at the same location get masked.

Another experiment with video tokens can be conducted to determine the importance of the temporal dimension of the video. Each video has 300 frames, which are divided by 30. The resulting 10 frames represent the tokens of the temporal dimension of the video. Analyzing the Shapley values of these tokens gives information about the model's use of this temporal dimension.

3.4.1.1.5 Prediction function MM-SHAP was originally tested for tasks such as VQA and image-sentence alignment, basing the Shapley value on discrete task scores. In the proposed method, the Shapley values are calculated based on the changes in softmax values to the predicted emotion label. In the experiments with running the modified

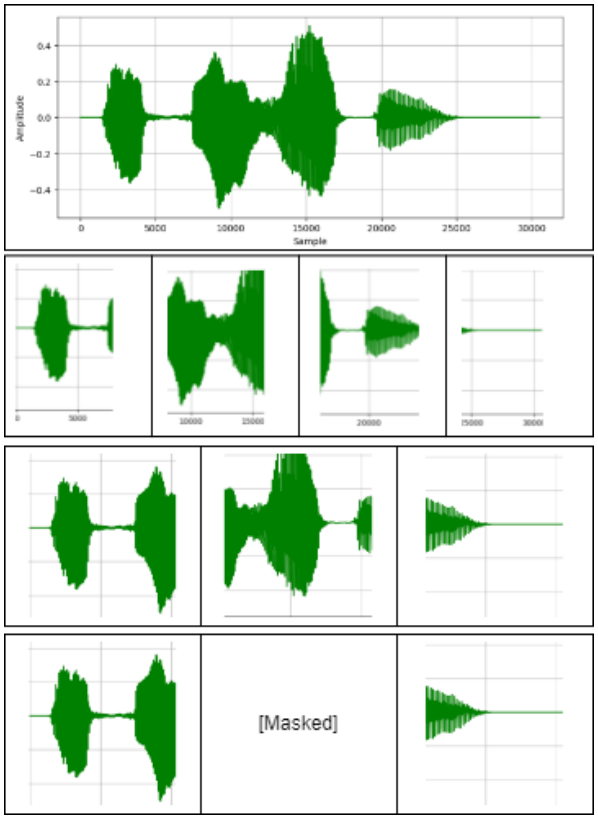


Figure 3.2: Example of the process of masking tokens for the audio modality.

Figure 3.3: Example of the process of masking tokens for the spatial dimension of the video modality. Each video is divided into a 4 x 4 grid. For each frame in the video, the patches at the same location get masked.

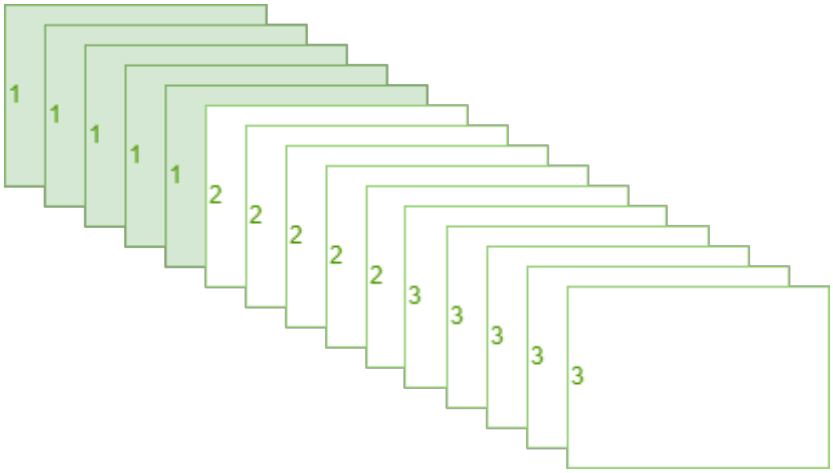


Figure 3.4: Example of the process of masking tokens for the temporal dimension of the video modality.

MM-SHAP on the MELD test set, Shapley values are calculated for each sample for each emotion class. For the implementation for the archive of Sound & Vision, the Shapley values are only calculated according to the predicted emotion class label. This approach, based on the calculation defined in Section 3.4.0.1, provides effective interpretability for the relevant class. Calculating Shapley values for all possible emotion classes can be computationally expensive, especially when dealing with larger datasets or real-time analysis. By focusing on the predicted class, the computational costs are reduced. The primary interest lies in understanding why a particular emotion was predicted by the model, in order to gain trust in the model. While a complete analysis of all possible emotions is important for the current research, it is often more practical to understand the decisions behind the predicted emotions. For example, if a researcher at the S & V is analyzing a video, they are likely more interested in the factors that led to the specific predicted emotion rather than the reasons why other emotions were not predicted.

3.5 Visualizations

The interpretability of the model's predictions and the Shapley values can be increased with the help of visualizations. The SHAP library[6] provides visualization plots for all kinds of analysis. A complete summary of the plots available can be found in Appendix A. These plots, however, are only suited for uni-modal models. Challenges come up when visualizing multimodal data, much like when choosing how to represent the data and processing it, as discussed in Section 2.3.1. Decisions that have been made in the process of representing the input as tokens have consequences for the interpretability of

[6] https://shap.readthedocs.io/en/latest/

the visualizations. For example, when masking the video input, tokens are represented by patches. The number and the size of these patches determine how detailed the explanation is. Smaller patches provide a detailed explanation but may hide global patterns, while larger patches offer a broader explanation but may overlook details. This is also the case for the number of tokens in the temporal dimension. These choices are directly reflected in the explanation.

In any case, visualizations can help make more informed assessments about the model's decision process and give transparency to the model's predictions. Visualizations highlight possible biases in the model and, therefore, can help evaluate its fairness. The designs of the visualizations for interpreting the Shapley values are given in Section 3.5.1 (text), Section 3.5.2 (audio), Section 3.5.3 (spatial dimension of video), and Section 3.5.4 (comparison between text and audio).

3.5.1 Visualizing Shapley values for text

For each individual sample, the interpretability framework can visualize a bar plot to explain the Shapley values given to the text tokens. As can be seen in figure 3.5, the y-axis lists features in descending order of importance to the model output, with the most influential features to the prediction at the top and the least influential at the bottom. Note that the order is based on the importance of the prediction of the sample and does not represent the overall importance of a feature. However, it is likely that features that have a high positive or negative contribution to the prediction of a sample, are influential in the general decision making process of the model. Recall that positive Shapley values contribute positively to the model prediction, and vice versa for negative values.

The x-axis represents the scale for Shapley values, with a vertical line at zero. Values to the right of the line are positive, while those to the left are negative.

Color is used to support the intuitiveness of the Shapley values. To increase contrast in the visualizations, positive values are displayed in red and negative values are displayed in blue. When explaining the influence of the text modality on the model's prediction, visualizing the individual words makes a good explanation, as humans know the meaning of the words and can judge whether the Shapley value assigned to them is fair to the model outcome.

Contextual information, such as adjacent words, syntax, and semantic relations, can influence the Shapley values too. Words can have different meanings and/or more nuanced sentiments depending on their context. The proposed visualizations for text are used to analyze unique words in a sentence in Section 4.2.2.1 and the contextual information in Section 4.2.2.2.

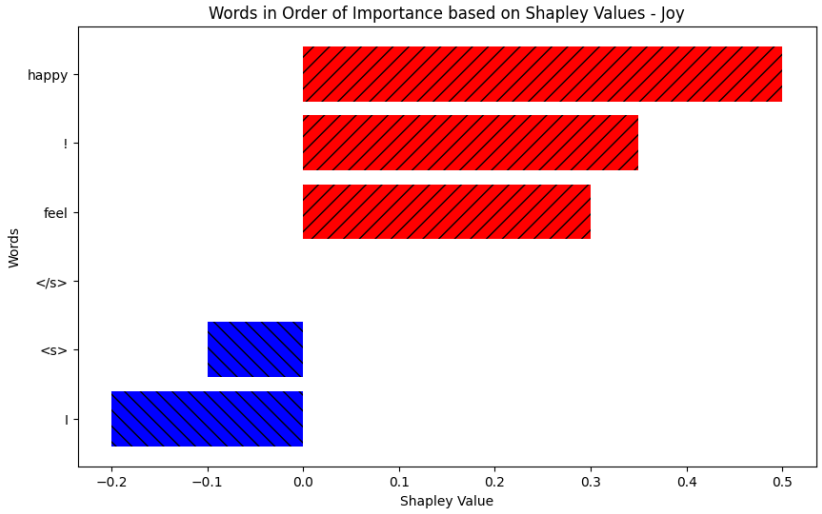


Figure 3.5: Visualization design for the mock Shapley values for the text 'I feel happy'.

3.5.2 Visualizing Shapley values for audio

The audio input for the model SSE-FT is represented as a vector with waveform values. As previously explained in Section 3.4.1.1.3, Shapley values for audio are obtained for each half a second of a waveform; the remainder of the waveform is appended to each segment, as can be seen in Figure 3.2. As each Shapley value represents a chunk of audio along the temporal dimension, it is important for the visualizations to correctly map these values to the right time stamps. Therefore, first the waveform is depicted with its amplitude along the y-axis and values, set to time in milliseconds, along the x-axis. Then, below the waveform, along the same x-axis, the Shapley values for each audio token are displayed along the y-axis.

Shapley values can be quite low, and they still need to be visible. For this reason, in the visualization, the audio Shapley values are shown normalized to the max and min values. The colors are blue (negative values) and red (positive values). An example of a visualization for the audio modality can be seen in Figure 3.6. In the example, the waveform for the utterance 'I feel happy' can be seen together with the example Shapley values for 3 tokens; -0.3, 0.5 and 0.7.

As can be seen in the visualization, the tokens do not completely align with the words in the utterance. The spikes in the amplitude correspond to syllables instead of words, and this requires some caution in the interpretation. Nevertheless, this design allows users, while listening to the audio, to verify whether these highlighted segments indeed contain emotional cues. This increases the interpretability and reliability of a model.

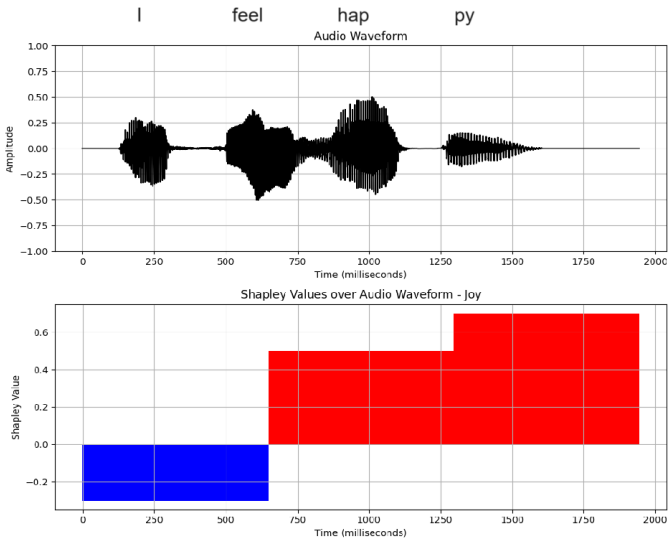


Figure 3.6: Visualization design for the mock Shapley values for audio 'I feel happy', the top graph shows the audio waveform over time, while the bottom graph shows the mock Shapley values repeated to match the total time length of the audio.

3.5.3 Visualizing Shapley values for video

The Shapley value visualization for the visual modality closely mimics the original image plot of the SHAP library shown in A.310. However, the proposed method makes use of superpixels instead of visualizing the Shapley values for each pixel. As discussed in Section 3.5, smaller patches, such as individual pixels used in the original SHAP image plot, allow for more details but require significantly more computational resources due to the higher number of patches processed. To balance interpretability with computational efficiency, the frame is divided into 16 superpixels on a 4x4 grid. This choice ensures sufficient detail in the visualization while reducing computational complexity in the process of running the interpretability framework [75].

In analyzing the visualizations shown in 3.7, the important question to the user is: Do the patches correctly point at the most emotional parts of the frame? The intensity of the colors, red for positive contribution and blue for negative contribution, indicates the contribution to the predicted emotion. By inspecting the Shapley value overlays, users can assess whether the model's focus aligns with their intuition of emotionally significant regions within the image.

Regions that are known to convey emotional information, such as faces, would be expected to be highlighted in the visualization. Other emotional cues could be body language, such as gestures or posture. For instance, hands might be highlighted in

various emotional states, or the entire upper body could indicate a state of defensiveness or confidence. Context within the frame also plays a role in conveying emotion. When multiple people are in a frame together, their proximity to each other can be an emotional cue, being closer together could indicate affection or aggression. Moreover, people, in context with other items, can show emotional states. It would be interesting to see if certain items in a certain context would also be highlighted as related to a certain emotion.

If the highlighted patches consistently fail to align with the regions that would be expected to have emotional information, this would raise a concern regarding the model's reliability or could suggest a bias in the training data. On the other hand, if the parts of the frame that are highlighted, intuitively align with human perceptions of emotional cues, this would suggest that the model effectively captures emotions.

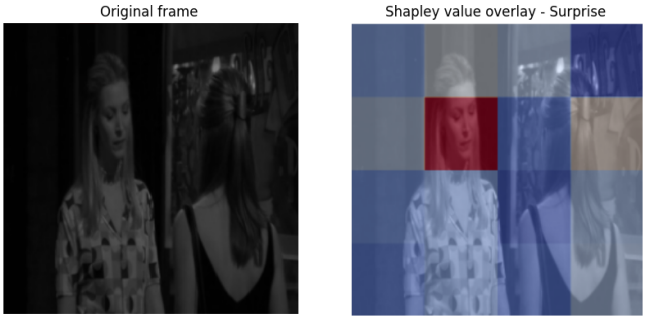


Figure 3.7: Visualization design for the mock Shapley values for the visual modality, the left subgraph shows the original frame in black and white, while the right subgraph presents the Shapley value overlay, in which the color intensity represents the contribution to the emotion label.

3.5.4 Visualizing the comparison between Shapley values for audio and text

By analyzing how the Shapley values of words align with the Shapley values of audio tokens within an utterance, insights into what information from the audio modality the model finds important for emotions can be gained. A visualization of the alignment between Shapley values for text and audio modalities needs to both show the words and incorporate the temporal dimension to correctly match both modalities.

Along the y-axis, the Shapley values for both audio and text are shown scaled to the maximum and minimum values. On the x-axis, the utterance is displayed. Each word's Shapley value is represented by a bar, where the color (as well as the position on the y-axis) indicates whether the contribution is positive (red) or negative (blue). The Shapley values of the audio tokens are connected with a line, the color of the line as well as the position on the y-axis reflect the overall Shapley value of the token.

In Figure 3.8, the design example of the comparison between Shapley values for audio and text can be seen of the utterance 'I feel happy, yesterday I did not'. The example Shapley values for text and audio are 0.1, 0.3, 0.5, 0, -0.3, -0.1, -0.2, -0.5, and 0.1, 0.4, 0.2, -0.5, respectively. When analyzing this visualization for emotional utterances, multiple observations could be made.

If the Shapley values of both modalities align, this would suggest that within the audio modality, the semantics, contribute most to predicting emotions. This implies a strong correspondence between the linguistic content and the acoustic features captured by the model.

If no alignment is observed between the modalities together with high audio Shapley values, this would indicate that other acoustic features beyond semantics might be more influential, i.e., pitch, intonation, or prosody. For instance, in the example in Figure 3.8, the third audio token has a positive Shapley value (0.2), but it corresponds to the word 'yesterday,' which has a negative Shapley value (-0.3). This lack of alignment suggests that the audio token might be carrying emotional cues through prosody or intonation that are not captured by the text alone. This could imply that the emotional tone conveyed by how 'yesterday' is spoken is significant for the model's prediction, despite the semantic meaning of the word itself not being positive in the given context. Furthermore, this visualization gives information about the contextual information in the text modality.

3.6 Implementation for the Institute of Sound & Vision

The current thesis is in collaboration with the Dutch Institute of Sound & Vision. As a product of the current thesis, the multimodal transformer SSE-FT is deployed in the institute's pipeline together with the proposed augmented MM-SHAP framework for interpretable emotion recognition. The implementation can handle raw video's with audio files as input from the archive of any size. The implementation first uses the Automatic Speech Recognition (ASR) model FasterWhisper[7] to segment the full audio and video file in short utterances. The speech from the segments is extracted and tokenized with RoBERTa (or with the Dutch textual backbone variant RobBERT). The output of the proposed system is the predicted emotion label for each video segment

[7] https://github.com/SYSTRAN/faster-whisper

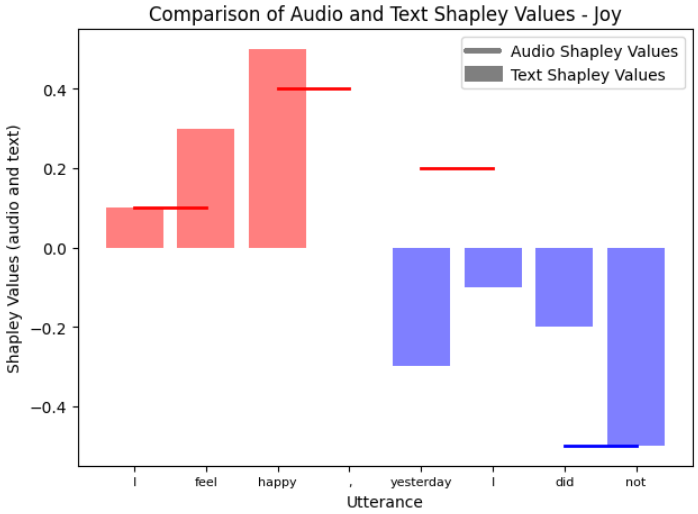


Figure 3.8: Visualization design for comparing the Shapley values of the audio and text modality, the graph shows both text and audio Shapley values together, with text Shapley values represented as bars for each word and audio Shapley values as lines over the corresponding segments of the text.

and the contribution values for each modality.

The contribution of the input within each modality is visualized according to the method described in Section 3.5. The visualizations of the Shapley values provide interpretability in the decision of the model for a segment. They can be used to analyze the characteristics in the video that contribute to emotion, for example, to study a particular cultural phenomenon or person. Moreover, in the context of producing a fair analysis for a data story, analyzing modality contributions can help identify biases that might exist in specific modalities, the model itself, or both.

3.6.1 The Dutch SSL model RobBERT

RoBERTa, the textual SSE model, has the ability to generalize well over other languages, however, the English model may have trouble capturing the specific nuances and cultural context of emotions expressed in Dutch. Therefore, in an effort to improve performance, RoBERTa has been replaced with RobBERT, a Dutch language model. The architecture and pre-training methods of RobBERT are based on the RoBERTa model, and the model has been pre-trained on OSCAR, a large Dutch corpus containing nearly 126 million lines of text [31].

The similarity of RobBERT to RoBERTa makes the swap to the Dutch model reliable since the architecture of the model has not drastically changed. The embedding dimensions of the textual baseline have been altered from 1024 to 768. The MELD

train and validation utterances were translated into Dutch using the 'googletrans' library from Google Translate[8]. The model was then trained on the translated data, keeping the audio and video modalities the same.

To check architectural compatibility between RobBERT and SSE-FT, RobBERT is trained and evaluated on the translated MELD utterances. To find out if improvement is possible by making balancing adjustments to the dataset, 75% of neutral samples was eliminated. In both experiments, RobBERT was trained for 5 epochs with batch size 16.

3.6.2 Sound & Vision case study

The proposed implementation of SSE-FT with the modified MM-SHAP is evaluated on a selection of videos annotated by Maddalena Ghiotto, who was an intern at Sound & Vision researching the ontology of annotations for MMER in media culture [43]. For her research, Ghiotto selected eight videos from diverse Dutch TV shows to form a corpus and filtered videos on two criteria:

1. Items more likely to portray people expressing emotions: to fullfil such a requirement, after a preliminary exploration of items across different genres, talk shows were identified as the most appropriate genre, as they mostly portray two or more people engaging in a conversation that involves exchanging opinions about a topic, often in a subjective and emotionally charged way.

2. Items about queer discourse, in order for Ghiotto to enable further topic analysis on queer archival material.

From this corpus, Giotto annotated segments for 4 videos that related to queer topics, as can be seen in Table 3.2. As previously discussed in Section 2.4.1, it is crucial for annotating emotions to have a clear understanding of the different ways emotions can be annotated, and make a clear approach for this. In her annotation process, Ghiotto only annotated the multimodal emotion conveyed by the speaker of an utterance, since this is the only case where all modalities refer to the same emotion. Each annotation includes a 'trigger,' which is a precisely identifiable part in one or more modalities (text, audio, video) where the annotator recognizes the expression of emotion. For example, a trigger could be a frowning facial expression in the video, the phrase 'I am hurt' in the text, or a louder pitch and tense articulation in the audio. Triggers are categorized as verbal, visual, or aural [43]. Ghiotto annotated each segment with the emotion category according to Ekman, the agent who conveys the emotion and the emotional triggers. Out of 145 utterances, 68 were found to be not neutral.

[8] https://pypi.org/project/googletrans/

Name of Program	Date	Broadcaster	Summary of Content
WELLES NIETES	19-03-1988	NCRV	Discussion about the statement: 'Can you say what you think on television?'. In the selected part, the conversation focuses on the representation of transgender people in the program De Nacht, featuring Adelheid Roosen. The discussion is moderated by Legien Kromkamp.
HET BLAUWE LICHT	17-03-1999	VPRO	Biweekly discussion program from De Balie in Amsterdam presented by Anil Ramdas and Stephan Sanders, in which various guests discuss topics and the way in which they are depicted using television fragments and photos. In this episode, the guests comment on some extracts from a documentary about a transgender person's transition (Vergezicht: Body and soul, RVU, 18/03/1999).
TIJDVERSCHIJNSELEN	31-03-1985	VPRO	Discussion program led by Ad 's-Gravesande, this episode features (mostly pregnant) women discussing various ways of having children, such as traditional methods, artificial insemination, and in vitro fertilization. Topics also include raising children in traditional families, single-parent families, and same-sex couples, as well as motherhood, surrogacy, and family planning.
SONJA OP ZATERDAG	06-04-1985	VARA	Saturday version of Sonja Barends' weekly talk show in which she talks to various guests about a current topic. In the selected part, the poet and translator Jim Stratton Holmes is invited to talk.

Table 3.2: Description of the annotated selected corpus from Ghiotto [43].

3.6.3 Case study evaluation

To analyze and evaluate the proposed implementation of SSE-FT with the modified MM-SHAP on the typical media from the S & V archive, a case study is conducted on the corpus with selected non-neutral annotations as discussed in Section 3.6.2. This involves analyzing the Shapley values for each sample and checking if the key features (words, audio tones, and visual cues) that contributed to the emotion prediction compare to Ghiotto's annotated emotional triggers. For each sample, the visualizations for the textual modality are analyzed to check if SSE-FT understands the emotional cues.

The utterances from this corpus are extracted from the ASR used in the 'Media Suite', the research platform used by Sound & Vision. It was noted that the utterances extracted are quite long, spanning multiple sentences and pauses. In contrast, SSE-FT is trained on smaller utterances, spanning between speaking pauses. The longer utterances may have more emotional information than the shorter ones, or they might be more complex. For this reason, a precise assessment in terms of accuracy can not be made, however, a sample-by-sample analysis is done by the author to evaluate the shorter utterances including the annotated emotion trigger words. For example, the utterance 'Nou ja ze begonnen dus mee met die bossen te schreeuwen en dergelijke op een vervelende manier ja ik ben doodzenuwachtig dat dat mag misschien wel maar ik vond het heel vervelend want er lopen een heleboel van die mensen rond ik heb zelf persoonlijk eentje gekend.', is annotated with 'Fear' due to the verbal trigger 'ik ben doodzenuwachtig'. To analyze the proposed method, the new utterance 'Ja ik ben doodzenuwachtig dat dat mag' is examined to see whether SSE-FT accurately predicts the emotion of 'Fear' for this segment and if the Shapley values highlight the verbal triggers.

Chapter 4

Results

In this section, the results of this research are presented. The results of finetuning SSE-FT on the MELD dataset are given in Section 4.1. The results from applying the modified interpretability framework MM-SHAP are discussed in Section 4.2, zooming in on the modality contribution in Section 4.2.1 and the textual modality in Section 4.2.2. Moreover, in Section 4.2.3, the performance of SSE-FT on the corpus of the Institute of Sound & Vision is discussed.

4.1 Finetuning SSE-FT

The finetuning method used for SSE-FT is described in Section 3.3.2. Upon evaluation of the SSE-FT model with all modalities on the test set, performance metrics were obtained and can be found in table 4.1. The accuracy on the test set with fp16 was 58,0%. Accuracy remained the same when fp16 was deactivated. In their study, Siriwardhana et al. reported a higher performance on the MELD dataset, specifically an accuracy of 64.3% [97]. However, it is not explicitly stated on which dataset this result was obtained. During the training process of this research, the model achieved its highest accuracy of 66.7% on the validation set at epoch 13. However, when the model was evaluated on the test set using the specified hyper-parameters (trained for 20 epochs), the obtained accuracy was lower.

Dual accuracy was calculated to assess the ability of the model to recognize emotions. Moreover, the precision, recall, and F1 scores for both the neutral class and the non-neutral class were calculated. Both results can be seen in Table 4.1. The accuracy for the neutral class and the non-neutral class is 89,0% and 45,1% respectively. Averaging this gives a dual accuracy of 67,0%. For the neutral class, the precision is extremely high, indicating that all predictions are correct (precision = 1). However, the lower recall for non-neutral classes (recall = 0.45) means that the model is missing many

actual instances of non-neutral emotions. The classification results can be seen in Figure 4.3. The model fails to predict the emotion labels: fear, sadness and disgust. In Figure 4.1, the confusion matrix shows the classification mistakes.

Table 4.1: Performance metrics of SSE-FT.

Metric	**Neutral Class**	**7-Class Accuracy**	**Non-neutral Class**
Overall Performance (Test Set)			
Accuracy	0.89	0.58	0.45
Precision	1.00	0.48	0.83
Recall	0.89	0.58	0.40
F1 Score	0.95	0.51	0.55
Reported by Shiriwadhi et al.			
Accuracy		0.64	

Table 4.2: Ablation study results of SSE-FT (Test set)

Modality	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Text	0.60	0.54	0.60	0.54
Speech	0.48	0.48	0.48	0.48
Video	0.48	0.48	0.48	0.48
Audio & Video	0.48	0.48	0.48	0.48

Table 4.3: The distribution of predicted emotion classes with the percentage to the total for each emotion class.

Class	Total	Predictions			
		Neutral	Anger	Joy	Suprise
Neutral	1256	1123 (89.43%)	2 (0.16%)	96 (7.65%)	35 (2.79%)
Sadness	208	147 (70.67%)	11 (5.29%)	40 (19.23%)	10 (4.81%)
Anger	345	105 (30.43%)	32 (9.28%)	155 (44.93%)	53 (15.36%)
Joy	402	134 (33.33%)	25 (6.22%)	229 (56.97%)	14 (3.48%)
Suprise	281	65 (23.13%)	24 (8.54%)	64 (22.78%)	128 (45.55%)
Fear	50	27 (54.00%)	1 (2.00%)	15 (30.00%)	7 (14.00%)
Disgust	68	32 (47.06%)	8 (11.76%)	21 (30.88%)	7 (10.29%)

4.1.1 Ablation study

After training the model SSE-FT for each individual modality, the results shown in Table 4.2 were obtained on the test set; the text modality had an accuracy of 60.0%, the other ablation studies resulted in an accuracy of 48.1%. From these ablation study

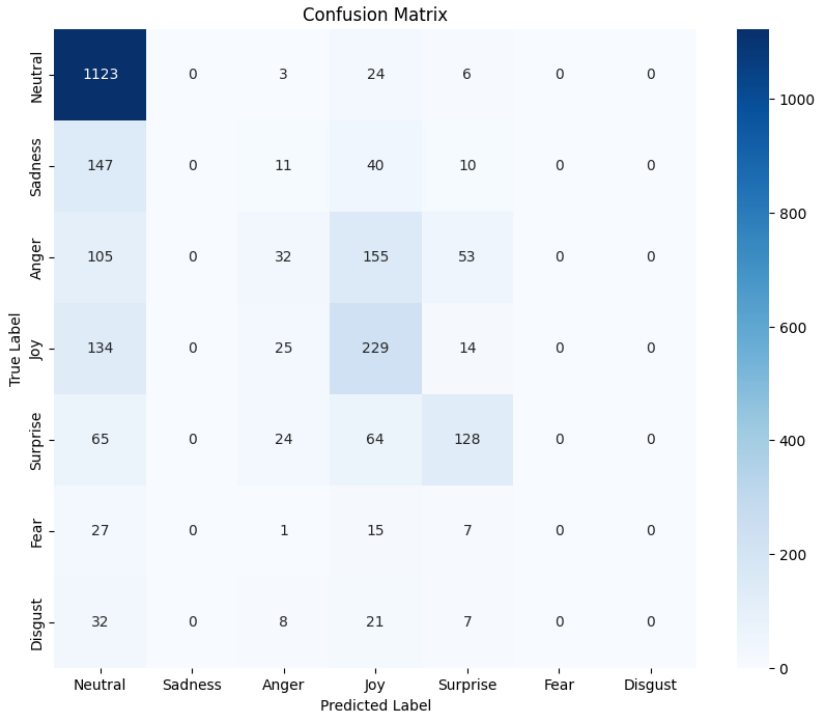


Figure 4.1: The confusion matrix illustrating the classification performance of SSE-FT across all emotion labels.

results, it is clear that the model prefers the text modality for predicting emotions from the MELD test set. To assess whether the audio and video modality can improve each other's performance without the text modality, an ablation study was conducted to assess the impact of combining the audio and video modalities on model performance. The resulting accuracy obtained from this combination was 48.1%. Without the text modality, all of the samples are predicted to have the neutral emotion label. In the test set, there are 1256 neutral samples out of 989 samples in total, this is 48.1% neutral samples, which explains the score.

4.2 Experiments with MM-SHAP

The proposed interpretability framework, explained in Section 3.4, offers a more comprehensive understanding of the model's decision-making process compared to the ablation studies. While the low performance of the uni-modal and dual-modal models may indicate issues regarding multimodality, this often doesn't provide specific insights into what exactly is going wrong and where the limitations in these models lie. Ap-

plying MM-SHAP provides a direct metric for which modality contributes the most to the model's output. Unlike performance-dependent metrics such as the ablation study, MM-SHAP offers insights into modality contributions regardless of the model's overall performance. This means that even for misclassified samples, MM-SHAP can identify which modality had the greatest influence on the model's decision. In an ablation study, the model is trained on a subset of modalities, for this reason, the metric may not account for the interaction between modalities in the full model. Conducting experiments with the full multimodal model intact can offer a more accurate assessment of its performance. The results regarding the modality contributions are discussed in Section 4.2.1, and the Shapley values from the textual modality are analyzed in Section 4.2.2.

While initially, this research included plans to analyze Shapley values for text, audio, and video, the current implementation of SSE-FT trained on the MELD dataset does not use the audio and video modality. It attributes near-zero Shapley values for the audio and video modalities due to uni-modal collapse. Hence, the analysis of the video and audio modalities is infeasible. Nevertheless, the visualizations for the audio and video modality as well as the comparison visualization for audio and text Shapley values are ready to use as they are described in Sections 3.5.2 (Audio), 3.5.3 (Video), and 3.5.4 (Text and Audio).

4.2.1 Modality contribution

Upon testing SSE-FT with all modalities on the MELD test set, it became apparent that computing the overall modality contribution on the complete MELD test set was too time-consuming for the available computational resources. As a result, a selection process was undertaken, which included correctly classified samples for each emotion class, as well as samples misclassified for each other emotion label. This selection process yielded a total of 55 samples for analysis. When assessing the overall modality contribution on this selected sample set, it was found that the textual modality achieved a contribution of T-SHAP of 99.8%. This indicates that SSE-FT likely experiences uni-modal collapse on the MELD dataset, where the model heavily relies on a single modality (in this case, text) while disregarding the information from other modalities.

4.2.2 Analyzing the textual modality

The interpretability framework outputs Shapley values for each textual input token. These Shapley values are helpful for understanding the importance of each part of the input, but they first need to be put into context for their correct interpretation. To increase interpretation, the Shapley values are visualized for each sample in the selected MELD test subset, as described in Section 3.5.1. In Section 4.2.2.1, the text modality

from the samples is analyzed, visualizing the unique words in a sentence, without their specific context. In Section 4.2.2.2, the text modality is analyzed with the contextual information.

4.2.2.1 Semantic influence

Within an utterance, a distinction can be made between words and paralinguistic cues. Each category has another role in carrying meaning and nuance, and thus in distinguishing emotional states.

4.2.2.1.1 Paralinguistic cues From the textual Shapley value visualizations, it becomes apparent that punctuation is the most important cue for emotion. In all samples that are predicted as 'Joy' , the token '!' was represented as least once, and has a total Shapley value of at least 0.25. An example of this can be seen in Figure 4.2.

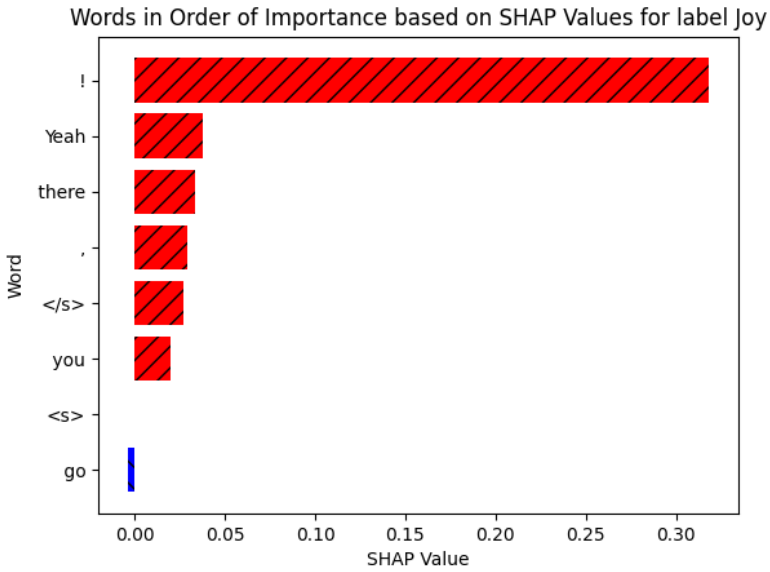


Figure 4.2: Visualization of the Shapley values for the utterance 'Yeah, there you go !'.

In the textual modality, punctuation can function similarly to prosody in audio, which includes aspects like volume, pitch, and intonation. Punctuation can be interpreted as paralinguistic cues that express the emotional state of the writer [22].

It's important to note that paralinguistic cues alone don't determine a writer's or speaker's emotion. Instead, a fuller understanding emerges from considering verbal meaning alongside these cues, within the context of the text's type and genre. In

samples that are predicted as 'Surprise', the tokens '!' and '?' as well as their alterations '!?' and '!!' are frequent and have high Shapley values. For the emotion surprise, these tokens are often co-occurring with the words 'what', 'Oh', 'Huh', 'No', 'Why', 'Who', and 'God'.

The results indicate that in the context of transformers trained to recognize emotional cues, paralinguistic cues from punctuation can give a strong signal that a sentence is not neutral.

4.2.2.1.2 Emotional words It would be expected to see certain emotional words highlighted with high Shapley values. However, it seems that a sample never gets predicted emotionally because of a single emotion baring word. In all cases, a punctuation emotion cue breaks the neutrality. There are examples where emotional words such as 'frown' or 'betrayal', 'sorry', 'broke', 'worry', and 'great' contribute negatively to a neutral label.

If a word contributes negatively to a neutral label, it makes sense that this word contribute positively to an emotional label, like 'sadness', 'happiness', or 'anger'. For example, it would be expected if the word 'sorry' would contribute positively to the label 'sadness'; this implies that the model has captured the semantic meaning of the word and accurately associates it with the emotional context of sadness. Some examples from the selected test are zoomed into.

1. The utterance 'I mean if you buy a bed from Janice's ex-husband, that's like betraying Chandler.', is annotated as anger. The word 'betraying', contributes -0.12 to the neutral label, +0.028 to the anger label, +0.032 to joy, and +0.008 to surprise. As can be seen in the confusion matrix, joy and anger get swapped often, and the model is not nuanced enough to differentiate between the two.

2. The utterance 'When I get up there I'm going to kick some ass' is annotated with anger. While 'When I get up there' remains neutral, 'I'm going to kick some ass' shifts the tone away from neutrality, namely 'kick' and 'ass' contribute -0,1 and -0.17 to the negative label respectively. The same words contribute positively to all other emotion labels. 'Kick' contributes 0.04 to joy and 0.016 . 'Ass' contributes 0.08 to joy and 0.016 to anger. The visualization of the contribution of the tokens from this utterance can be seen in Figure 4.3.

4.2.2.1.3 Neutral bias In the example in 4.3, we see negative Shapley values for the majority of the tokens in the sentence; however, such as in this sample, in some cases the sample is still predicted as 'Neutral'. When looking at the base value for the neutral label for this sample, it is observed to be 0.832, suggesting a high baseline prediction for neutrality. This is supported by the observation that Shapley values for tokens

indicating positive contributions to emotional labels are significantly lower than those indicating negative contributions to the neutral label.

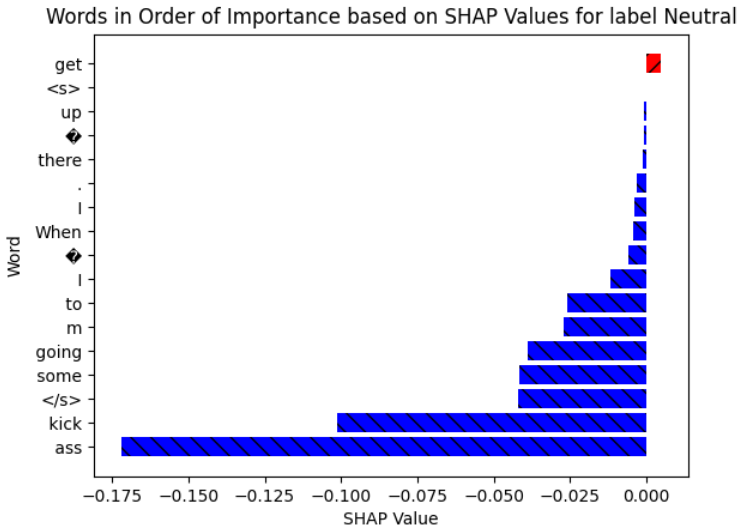


Figure 4.3: Visualization of the Shapley values for the utterance 'When I get up there I'm going to kick some ass'.

4.2.2.2 Contextual influence

Does the model use context for emotion prediction? In Figure 4.3, it can be seen that fairly neutral words like 'I', 'am', and 'going' contribute negatively to the neutral label and positively to other emotion labels. In the context of the sentence 'I'm going to kick some ass,' the fairly neutral words gain emotional weight and contribute to the overall sentiment. Another example is the utterance 'I broke it.' as shown in Figure 4.4. The word 'it' is inherently neutral; however, when combined with the word 'broke,' it acquires a non-neutral weight. The phrase 'broke it' reveals that 'broke' implies physical damage, thereby conveying a negative sentiment. All other meanings of the word 'broke' convey a negative sentiment: failure or bad financial status. The word 'it' specifies the object affected by the negative action. These examples suggest that the model can understand context and uses it to detect the emotional tone of a sentence.

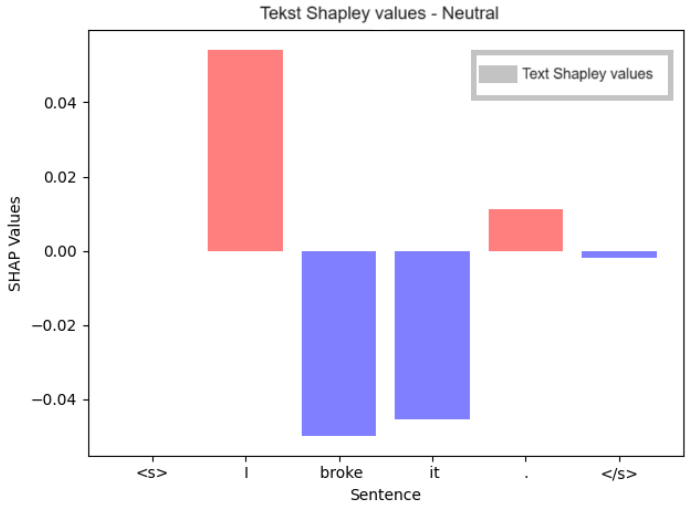


Figure 4.4: Visualization of the Shapley values for the utterance 'I broke it.'.

4.2.3 Evaluating SSE-FT on a selected corpus from the Sound & Vision archive

The model SSE-FT with the interpretability framework was evaluated on a selected corpus from the archive of Sound & Vision described in Section 3.6. The model consistently misclassified emotional utterances, assigning a neutral label to all instances. This outcome was anticipated, as the model relies only on the textual backbone trained on English text data while the annotated utterances are in Dutch. RoBERTa, the textual SSE model, possibly cannot capture the emotional cues from the Dutch text. Therefore, in an effort to improve performance on the archive, RoBERTa has been replaced with RobBERT, a Dutch language model. This resulted in an accuracy of 48% on the MELD validation set with translated utterances.

The new instance of SSE-FT with the Dutch textual backbone again consistently predicted the neutral label on the corpus from the archive. A high 'neutral' baseline was observed: 0.49. Unfortunately, the annotated emotional triggers from the selected corpus described in Section 3.6.2 could not be compared to the Shapley values for audio, text, and video. Namely, running SSE-FT using RobBERT with the proposed interpretability model resulted in near zero Shapley values for all modalities.

To check whether the low performance is due to the architectural incompatibility or to RobBERT's general performance on the MELD dataset, RobBERT is trained and evaluated on the translated MELD utterances. This also results in an accuracy of 48%, in which only the neutral label is predicted. After deleting 75% of 'Neutral' samples, the accuracy on the validation set is 56%.

Chapter 5

Discussion

The aim of this research was to explore the interpretability of multimodal models for emotion recognition and to implement a multimodal model for the Sound & Vision archive. For this research, the state-of-the-art multimodal model SSE-FT was fine-tuned on the MELD dataset for emotion recognition. The performance of SSE-FT with all modalities included, measured with 7-class and dual class accuracy, was 58% and 67% on the MELD test set, respectively.

The interpretability framework MM-SHAP was used to investigate the multimodality of the model as well as to zoom in on each modality and find out what parts of the input are influential in recognizing emotions. MM-SHAP, tailored for vision & language models, was extended to handle the input format of SSE-FT, including text, audio, and video.

MM-SHAP was run on a selected dataset from the MELD test set representing correctly classified and misclassified samples, this resulted in T-SHAP 90.8%, I-SHAP 0.01% and A-SHAP 0.01%. These results indicate that the model SSE-FT has uni-modal collapse on the MELD dataset. To verify this result, ablation studies were done on the single modalities as well as with the combination of audio and video. These ablation studies showed that the uni-modal performance for text was 60.0%, and for each audio, video and audio-video 48.1% only predicting the 'neutral' class. These results confirm the uni-modal collapse. Part of this research included analyzing the influence of the input within single modalities on recognizing emotions. Due to SSE-FT only using the information from the text modality, the interpretability results for text from MM-SHAP are analyzed. The analysis of the visualizations for the textual modality supports the belief that the model has captured semantic and contextual emotional cues from the data.

For the implementation for the Institute of Sound & Vision, SSE-FT was trained on the translated MELD train set, substituting the English textual backbone roBERTa for the Dutch language model robBERT. SSE-FT was then tested on a selected corpus

from the Sound & Vision archive. In the following subsections, the results from SSE-FT on the MELD dataset (Section 5.1), the results from analyzing the contribution of the textual modality (Section 5.2), the results from testing SSE-FT on the Sound & Vision archive (Section 5.3) are discussed. Furthermore, the limitations of this research are given in Section 5.4, and future research is described in Section 5.5.

5.1 The performance of SSE-FT on the MELD dataset

In this section, the performance of SSE-FT on the MELD dataset is discussed. First, the results from the current research in comparison to the reported results in the original paper of SSE-FT is discussed in Section 5.1.1. Secondly, the causes for SSE-FT's low performance and its uni-modal collapse are explained in Section 5.1.2. And lastly, SSE-FT's ability to discern emotions is discussed in Section 5.1.3.

5.1.1 Reported results in the original SSE-FT paper

As described before, SSE-FT performs significantly lower on the MELD test set than reported in the original SSE-FT paper, namely 58% instead of the reported 64,3%, while maintaining the same hyper-parameters. Apart from evaluating on the MELD dataset, the authors evaluated SSE-FT on the CMU-MOSEI dataset as well as the IEMOCAP dataset measured with averaged F1 scores, 87.0 and 84.2 respectively. MELD is the only dataset for which F1 scores are omitted. It is not clear what causes the performance difference, but it does raise an issue for reproducability. Other multimodal models bench marked on the MELD dataset obtained a similar result, namely the hierarchical biLSTM 60,8 % and QIN 61,9 % [119, 80]. None of the multimodal models benchmarked on the MELD dataset including SSE-FT, have performed an ablation study to assess multimodality.

5.1.2 Low performance and the causes for uni-modal collapse

SSE-FT has low performance on the MELD dataset. Ablation studies on the CMU-MOSEI dataset imply increased performance with the current multimodal architecture of SSE-FT. However the ablation study in this research shows that the uni-modal model for text outperforms the multimodal model with text, audio, and video. The results from applying the interpretability framework and the ablation studies indicate that SSE-FT only relies on the textual modality for prediction, resulting in uni-modal collapse.

The model could fall back on the textual modality possibly for a series of reasons. From the results reported, and based on the results of this research it is likely that the

uni-modal collapse has to do with SSE-FT not being able to fully use the visual and aural information from the training samples.

Firstly, it could be, when finetuning data is limited or imbalanced, that the strengths of the three self-supervised learning (SSL) models, RoBERTa, Wav2Vec2, and Fabnet, vary significantly in zero-shot emotion recognition tasks, and therefore the model might heavily rely on one of the pre-trained models. RoBERTa has robust language understanding capabilities and performs well in zero-shot emotion recognition, as has been tested on the MELD and IEMOCAP dataset [54]. Moreover, research shows that the frozen Wav2Vec2 model performs well in speech emotion recognition evaluated on the IEMOCAP dataset with minimal finetuning [111]. However, at Spoken Language Understanding (SLU), including Intent Classification (IC) and Slot Filling (SF), Wav2Vec2 performs poorly, which suggests that the frozen model cannot hold complete semantic information. No research is found on evaluating Fab-Net on the task of emotion recognition.

However, the exact architecture of SSE-FT was implemented (including the same embedding sizes) and evaluated on the RAVDESS dataset, and achieved the weighted F1 score over all emotion class labels of 81.68 for the multimodal model [24]. The uni-modal models achieved the weighted F1 score of 80.87 (Wav2Vec2), 82.87 (Fabnet), and 81.12 (RoBERTa). This indicates that the SSL models perform well for the task of emotion recognition, and the uni-modal collapse is not likely due to the architecture of the model.

The RAVDESS dataset contains 7356 recordings, including both speech and song, performed by 24 actors (12 male, 12 female) in North American English. The dataset consists of raw .wav and .mp4 files that cover the emotional states: calm, sad, happy, neutral, surprised, disgust, fearful, and angry as can be seen in Figure 5.1.

The performance difference of the proposed architecture on the MELD dataset compared to RAVDESS can be explained by several factors. MELD contains acted conversations from the TV show 'Friends,' with flat, context-dependent emotional expressions. In contrast, RAVDESS has controlled, scripted recordings with exaggerated emotional expressions by actors, making the emotions clearer and more consistent. Additionally, MELD has an imbalanced distribution of emotion classes. Out of the 9988 MELD training samples, 48.12% is neutral, and emotional labels are underrepresented; 15.4 % joy, 13.22 % anger, 10,77% surprise, 7.97% sadness, 2.71% and 1.91% fear. The imbalance is causing the model to overfit on the neutral samples, preventing it from learning the data distribution, which results in the model failing to correctly distinguish boundaries between different emotion classes. On the contrary, the RAVDESS is balanced over all emotion classes.

These differences might account for the uni-modal collapse to the textual modality. The textual emotional cues are stronger and more straightforward, since in text, emotions are directly expressed through words, phrases, and paralinguistic cues. In

contrast, the emotional cues in video and audio are much more nuanced and indirect, hence the model needs more or clearer data samples for each emotion to fully capture the nuances. MELD does not have enough clear samples from each emotion class for SSE-FT to learn these nuances in the audio and video modality, the model falls back to the more direct textual emotion cues for prediction. Imbalance is a frequent and natural occurrence in multimodal datasets for emotion recognition [71]. To solve the imbalance in the data, multiple methods have been proposed, and this issue is discussed in Section 5.5.

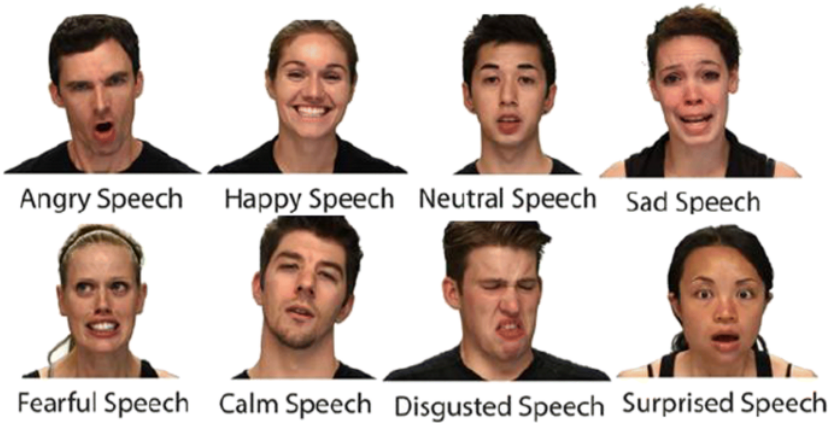


Figure 5.1: Examples of the eight RAVDESS emotions, Figure from [15]
.

5.1.3 Emotion class confusion and the need for interpretability

SSE-FT has trouble discerning the emotional class boundaries, however, when calculating performance with dual accuracy, which is 67%, it can be confirmed that in a lot of cases, the model does recognize the presence of emotion. The confusion matrix in Figure 4.1 shows that, apart from each emotion class to be confused with the 'neutral' class, emotion classes are often confused and classified as 'Joy' and 'Anger'. From the performance based metrics alone it is unclear what causes these confusions. This emphasizes the importance of creating interpretability, since these emotion class confusions among more details of feature importance are analyzed with the interpretability framework and discussed in Section 5.2.

5.2 Analyzing SSE-FT with the interpretability framework

For the exploration of the decision making process and performance of a multimodal model for emotion recognition, in this research, MM-SHAP is modified and applied

to SSE-FT. In this section, the representation of the modalities (Section 5.2.1), the effectiveness of the interpretability framework in assessing multimodality (Section 5.2.2) and the results from analyzing the textual modality with the interpretability framework (Section 5.2.3) are discussed.

5.2.1 Token representation

For MM-SHAP, representation tokens were created for each modality, and each modality is represented differently. In the following sections, the token representations for text (Section 5.2.1.1), video (Section 5.2.1.2) and audio (Section 5.2.1.3) are discussed.

5.2.1.1 Text token representation

The text modality representation is the most straightforward, since the textual input corresponds one to one with the tokens used as input to represent the textual modality for MM-SHAP. The model SSE-FT requires the text to be preprocessed with tokenization, without removing punctuation and start and end symbols. It is later determined that it is these paralinguistic cues, such as punctuation, that the model uses most in recognizing emotion. Removing these paralinguistic cues in the preprocessing phase might potentially prevent the model from overfitting on these types of emotional cues and prioritize semantic cues from words. Thereby, the model might increase its capacity to generalize on the textual input. However, training the model on a more balanced dataset would eliminate the need to apply this preprocessing step, since the overall capacity of the model to recognize emotions from text would possible be better.

5.2.1.2 Video token representation

For representing the video modality, the input of 256 x 256 is divided in 16 patches, so called 'superpixels' of the size 16 x 16. This method is cost efficient, since the number of tokens has a limit. For the task of emotion recognition, it might be more preferable if the different facial attributes, such as the spacing between facial features could be differentiated in the visualization. However, to acquire these details, the computational cost would be infeasible. For application, such as to the archive of the Institute of Sound & Vision, it is preferable if the interpretability method takes minimal time recourse. With the current method, the faces of the people in the videos from the MELD dataset often fit in two superpixels. Although the influence of the features in the face is less easy to interpret, if the model assigns a high contribution to the superpixels containing the face for a certain emotion this gives confidence in the outcome. Moreover, as previously discussed, the superpixels could be used to highlight other potentially emotional cues, such as a person's posture or items. For this reason, the proposed token representation for the visual modality is helpful for assessing the model's robustness.

A possible alternative to superpixels would be to create more semantically meaningful tokens. In their work, Cafagna et al. proposed such a method, using the visual concepts from the visual backbone of a Vision & Language model [16] which preserve semantics. However, their method is unsupervised and produces partitions that do not always exactly sum up to the total size of the image. To solve this issue, the authors create a leftover mask to fill in the unassigned space in the image, however, some tokens still overlap. This method is efficient for creating meaningful interpretations, although it creates an issue for the calculation of the modality contribution. For calculating T-SHAP, V-SHAP, and A-SHAP, the tokens can not overlap, as the visual modality would be attributed to much contribution. Possible solutions to this issue are discussed in Section 5.5.

5.2.1.3 Audio token representation

As previously discussed in Section 2.2.1.2, the audio modality is very important for discerning emotions, making it essential to include it in the analysis of emotions with MM-SHAP. The proposed modification of MM-SHAP is the first to incorporate the audio modality into the interpretability analysis of a multimodal model. In the proposed method, the tokens representing the audio modality are chunks of waveform values spanning half a second each. This method makes it possible to analyze the semantic properties in the audio and compare the contribution of a spoken word in the audio modality to that of the corresponding word in the text modality. The duration of half a second is an estimation of the average duration of a word in the MELD dataset. Limitations and future work regarding the token representation of the audio modality can be read in Sections 5.4 and 5.5.

5.2.2 The interpretability framework and uni-modal collapse

The proposed interpretability framework shows to be very effective in application. The uni-modal collapse is clearly visible in the results obtained from the interpretability framework, where T-SHAP is observed 99.8% for the selected samples. The interpretability framework eliminates the need for ablation studies on multimodality, as its results directly reflect the importance of each modality on the whole dataset. Ablation studies isolate one modality and might, because of this, overlook the interactions between the modalities in the complete model. Furthermore, as previously mentioned, MM-SHAP includes incorrectly classified samples in the multimodality assessment, and therefore creates a more exact measurement. MM-SHAP was originally tested for tasks such as VQA and image-sentence alignment basing the Shapley value on discrete task scores. In the proposed method, the Shapley values are calculated based on the changes in softmax values to the predicted emotion label. This research has demonstrated MM-

SHAP's ability to translate effectively to the task of emotion recognition with the proposed modifications.

5.2.3 Interpretability within the text modality

Apart from effectively assessing the multimodality of the model SSE-FT, the proposed interpretability framework also has the ability to provide modality specific interpretations and visualize these. Unfortunately, due to the uni-modal collapse, there are only results for the textual modality. The interpretability framework gives insight in how different tokens in the text modality are important for predicting emotions. The results show that paralinguistic cues such as punctuation attribute the most to the prediction of an emotional label. The emotion labels, 'joy', 'anger', and 'surprise' can be recognized by punctuation. Surprise is classified when the symbol '?' is present, 'joy' and 'anger' are often confused since both samples often have a '!' present in the utterance. From just the confusion matrix, it would be a guess as to why 'joy' and 'anger' are confused often, however, the interpretability gives us information on the process behind this prediction.

Furthermore, it was found, that in samples misclassified as the neutral label, a negative contribution is assigned to words that have an inherently negative or positive sentiment. This tells us that SSE-FT has captured some of the semantic meaning of the words with their emotional load. Moreover, apart from the semantic meaning of the tokens in the textual modality, the context of the words and their place in the sentence also influence their emotional load. The results suggest that the model uses context for predicting emotions. Emotionally neutral words receive a negative contribution to the neutral label when paired with other words that make the sentiment of the grouped words non-neutral. To make definite statements about the use of context by SSE-FT, further assessment is needed and is discussed in Section 5.5.3.

Additionally, a fairly high baseline for the 'neutral' label was observed from the Shapley values of the textual modality. As most samples are predicted 'neutral' and the base-line reflects the average prediction of the model, this is not unexpected. However, it is a clear sign that SSE-FT has a bias for the 'neutral' label, caused by the imbalance in the training data. This also reflects a lack of strong indicators for other emotional labels in the textual modality. This is supported by the observation that Shapley values for tokens indicating positive contributions to emotional labels are significantly lower than those indicating negative contributions to the neutral label. The lower Shapley values for tokens indicating positive contributions to emotional labels imply that the model may require stronger evidence from these tokens to decide which emotional label to predict, since now the contributions are divided over all emotional labels. The bias towards the 'neutral' label can be reduced, for instance, by applying class weighting or data augmentation, which is discussed in Section 5.5.

5.3 The performance of SSE-FT on the Sound & Vision archive

Due to the unavailability of well annotated Dutch datasets for multimodal emotion recognition, a cross-lingual approach was proposed, training on the American English MELD dataset and testing on Dutch videos from the Sound & Vision archive. From the results, however, it was shown that the MELD dataset provides insufficient emotional cues for the audio and video modality, and SSE-FT relied most on the paralinguistic text cues for discerning emotion. The English model was unable to recognize emotional content from the videos in the archive as paralinguistic text cues were not present in these videos.

For this reason in efforts to improve the performance of SSE-FT on the S&V, the model was trained and evaluated with the textual backbone RobBERT on the translated MELD utterances; however, this made performance worse. Evaluating SSE-FT with RobBERT as expected also resulted in poor performance on the selected dataset from the archive of S%V, only predicting the neutral label. The interpretability framework gave near-zero Shapley values and a high neutral baseline, which indicates that learned emotional cues are weak.

Research on the perseverance of emotion classes after translation from English to Finnish, French, and Italian deemed the degree of perseverance sufficient for cross-lingual approaches [74]. Difficulties arise when translation becomes ambiguous or incomplete. After carefully assessing the translations, utterances with English 'Slang' and typical phrases became somewhat ambiguous due to a lack of correct context. However, no radical changes in semantics were found, and paralinguistic cues were preserved. Because of time constraints, the proposed interpretability framework could not be applied to the model with the Dutch textual backbone, however, the performance suggests that the emotional cues in the translated text were too nuanced or ambiguous for RobBERT, and although preserved in the translations, the paralinguistic emotional cues this time were not sufficiently learned. Moreover, since only the text modality was translated in this model instance, the audio is still in English, the poor performance of the model with RobBERT could be caused by the incompatibility of the Dutch text with English audio. However, this is most likely not the cause since the performance of the English uni-modal model was at least as good as the English multimodal model. RobBERT was evaluated without the structure of SSE-FT to find out if the low performance was caused by implementation faults or architectural incompatibility, this resulted in the same accuracy as using RobBERT within SSE-FT. Eliminating 75% of 'neutral' samples in the train and validation set resulted in an accuracy of 56%, which suggests that solving the imbalance of the dataset can improve the performance of RobBERT as well as SSE-FT with RobBERT as the textual backbone.

5.4 Limitations

In this research, the interpretability framework MM-SHAP is modified to analyze the audio and video modalities. While, the results show that the proposed method correctly processes the model input and applies the SHAP Explainer to the multimodal model for emotion recognition, some limitations of this research have been identified and discussed. Limitations in testing the interpretability framework (Section 5.4.1), the representation of the modalities (Section 5.4.2), the resources provided by Sound & Vision (Section 5.4.3) and the MELD dataset (Section 5.4.4) are explained.

5.4.1 Limitations of evaluating the interpretability framework

Due to the uni-modal collapse of SSE-FT fine-tuned on the MELD dataset, only the textual modality could be explored. The method for incorporating audio and video is fully developed, but has not been used to analyze emotional cues as done with the textual modality. Therefore, the effectiveness of the proposed interpretability framework for the audio and video modalities has only been reasoned theoretically and has not been empirically tested.

This limitation could lead to potential issues in applications, as the framework's capability to provide interpretability for the video and audio modality through visualization is unverified. Additionally, due to the limited computational resources available, the proposed method was only evaluated on a subset of the MELD test set described in 4.2.1, which might affect the robustness of the evaluation.

5.4.2 Limitation within the interpretability framework

Some limitations related to interpretability lay in the choices made for representing the audio and video modality. In visualizing the audio modality, the amplitude is plotted together with the audio tokens. In this visualization, a user is expected to interpret the audio waveform. However, as spikes in the waveform represent syllables rather than full words, reading the waveform can become challenging, especially for longer words and utterances. This visualization is therefore more valuable when analyzing the emotional cues in the amplitude, but for other analyses, such as examining the semantics within the audio modality, this visualization might be less interpretable. In such cases, the visualization comparing the audio and text modality is preferred, as it has a clearer alignment between spoken words and their corresponding Shapley values.

Moreover, the current representation of the audio modality only considers the amplitude, ignoring other acoustic properties. Apart from speech, other sounds such as laughter, music, or ambient noises (e.g., rain, objects falling) can also convey emotions. These non-speech audio elements pose a challenge for the current method, since

they also spike the amplitude. While the method allows users to listen to the audio and verify whether highlighted segments contain emotional cues, it does not inherently differentiate between types of audio. It is worth noting, however, that in the MELD dataset, utterances are very short and mostly consist of spoken words.

To create interpretability in the semantic properties of the audio, a visualization of the comparison between audio and textual Shapley values has been designed. A limitation of this is that the audio and text tokens do not perfectly align with each other. There are often fewer audio tokens than text tokens, and an audio token is not designed to represent a single word. This misalignment limits the accuracy of the comparison and the interpretability of the visualization. To increase interpretability, a more exact estimation for the length of the audio tokens could be made, which is discussed in 5.5.

5.4.3 Limitation in the resources of Sound & Vision

Limitations can be identified in the application of SSE-FT fine-tuned on the MELD dataset for the Sound & Vision archive. The multimodal model SSE-FT for emotion recognition with the interpretability framework has been fully developed to function on the archive, however, due to SSE-FT's inability to fully use the information from the visual and audio modality, performance is quite low. The current instance of SSE-FT relies fully on the textual modality and is therefore effectively a uni-modal text model for emotion recognition.

Furthermore, during the process of this research, there was no dataset available for finetuning SSE-FT on data from the archive. Therefore, it was decided to use the large benchmarked dataset, MELD, for finetuning. Due to the unavailability of computational power from Sound & Vision, it was not feasible to finetune SSE-FT on data from their archive. The current implementation can easily be fine-tuned on a more representative dataset for the archive, whenever data and computational power become available.

5.4.4 Limitations in the MELD dataset

As previously discussed, issues regarding the uni-modal collapse arise due to limitations in finetuning SSE-FT the MELD dataset. Apart from the imbalance in training samples, several characteristics of the dataset could pose challenges. The MELD dataset consists of acted dialogues, which can differ a lot from the content in a television archive, such as talk-shows and interviews, showing conversations between real people. Acted emotions might be exaggerated and less nuanced compared to real emotions, affecting the model's ability to generalize well.

Moreover, the filming style and editing techniques in MELD, designed for comedic

timing, differ from the more static filming in television archive formats such as talk-shows. Additionally, the presence of a laughing track in MELD, added for comedic effect, may give artificial emotional cues that are not present in natural conversations. During the annotation process, some utterances with subtle emotional nuances may be perceived as more humorous due to the laughing track, which could bias towards annotating an emotion.

Furthermore, the MELD dataset mostly features the main seven characters from the Friends series in the conversations, while the archive has a wide variety of speakers. The model might learn specific patterns related to the seven characters, and this can affect generalization.

5.5 Future work

From the described limitations in Section 5.4, a series of directions for future research can be followed. The future work is structured as follows: solutions for increasing the performance of SSE-FT will be given in Section 5.5.1, improvements for representing and visualizing the modalities for the interpretability framework will be given in Section 5.5.2, additional experiments with the interpretability framework will be explored in Section 5.5.3 and lastly, a method, using the attention mechanism in transformers, for assessing the robustness of the proposed interpretability framework is discussed in Section 5.5.4.

5.5.1 Increasing the performance of SSE-FT for the Sound & Vision archive

As previously discussed in Section 5.1.2, the low performance of SSE-FT and the uni-modal collapse are most likely due to the inability of SSE-FT to capture visual and aural emotional cues from the MELD samples. This stands in contrast to its good performance on the RAVDESS dataset. The MELD dataset was chosen for this research because it includes multi-party conversations, that closely resemble the content of Sound & Vision.

To improve SSE-FT's performance for Sound & Vision, multiple approaches could be evaluated in future work. Initially, SSE-FT could be trained on the RAVDESS dataset, which contains clearer emotional expressions in the audio and video modality. This initial training could potentially result in good performance on the archive. Subsequently, SSE-FT could be finetuned on the MELD dataset or a specially crafted corpus directly from the archive.

Alternatively. another approach would be to address the performance issues of SSE-FT on the MELD dataset directly, namely, SSE-FT would require more clear emotional samples to capture visual and aural emotional cues. This can be acquired by applying

several approaches to fix the imbalance and neutral bias in the dataset. These approaches fall into three main research directions: data augmentation, sampling strategies, and loss-sensitive methods.

With data augmentation, new samples could be generated by modifying the original MELD samples for the underrepresented emotion class samples. Since the samples in the MELD dataset are multimodal, involving text, audio, and video data, this process becomes more complex, since most augmenting techniques are modality specific. One solution, that could be used for multimodal input, would be to create adversarial examples with a Generative Adversarial Network (GAN) [71].

A more straightforward way of fixing the imbalance is by applying a sampling strategy. Future research could balance the class distribution by doubling the samples of underrepresented emotion classes or halving the 'Neutral' samples. Eliminating 75% of 'neutral' samples, improved the performance of RobBERT on translated MELD utterances by +8% accuracy. Furthermore, imbalance could be tackled by applying a loss sensitivity method, by assigning higher weights to the loss function for underrepresented classes. For example, the loss of each sample could be multiplied by a weight factor that is inversely proportional to the sample class frequency. This way, errors made in predicting emotion classes such as 'Disgust' and 'Fear' have a greater impact on the overall loss, and the model learns to pay more attention to these emotions.

Moreover, as a cross-linguistic approach using the English RoBERTa and an English dataset may give lower performance on the Dutch video's, future works could use the multi-lingual version of BERT, mBERT, as a tokenizer and textual backbone for finetuning on both an English and Dutch dataset [34].

5.5.2 Improving the interpretability framework

Within the proposed interpretability method, a few improvements could be explored in future work. Firstly, the token representation for the spatial dimension of the visual modality could be more semantically meaningful, as discussed in Section 5.2.1.2. Future research can apply approaches such as those proposed by Cafagna et al., using the visual priors in the model. However, a method needs to be created to ensure a complete partition of the frame. This could possibly be done by defining the overlap and cutting the intersection from the token space. Other visual hand-picked masking techniques could be explored to experiment with visual emotion cues, such as partitioning the frame in front and background. Moreover, other object recognition methods could be explored to assess the emotional cues of items, faces, and body parts.

Secondly, to assess the ability of multimodal models to use textual context to discern emotion classes, NLP techniques can be used. For example, n-grams could be used to represent tokens, comparing Shapley values for different n-grams can give insights into the use of context. For instance, the bigram 'very happy' might have a higher Shapley

value than the individual words 'very' and 'happy' alone, indicating that the context provided by the combination is significant for emotion recognition. Moreover, adding 'not', creating the trigram 'not very happy' might provide context that changes the contribution of the token from positive to negative to the class 'Joy'. Additionally, Part Of Speech tagging could be applied to analyze syntactic patterns and context. For instance, a certain word can provide different emotional clues as a adverb compared to an adjective.

Lastly, future research can improve the alignment of the text and audio tokens in the Shapley value comparison visualization. By including time stamps at the word level, the framework could determine the exact length of the audio tokens to correctly assess the semantic properties of the audio modality. This would increase the interpretability for audio modality and the reliability of both the audio and the textual modality. Time stamps can be saved using an ASR method such as Whisper.

5.5.3 Additional experiments with the interpretability framework

Several other experiments could be suggested for future work with the proposed interpretability method. Firstly, once SSE-FT's performance in recognizing emotions is improved or if another multimodal model with good performance is available, the proposed interpretability framework should be applied to assess the video and audio modality. The following experiments mentioned in Section 3 can be conducted in future research. T-SHAP, V-SHAP, and A-SHAP can be calculated for each emotion label to get insights into how each modality contributes to the recognition of emotions. The results from these experiments can be compared to the results from Khalene et al., who applied SHAP to find out the influence of features extracted from the CMU-MOSEI dataset and corresponding modality on various emotion classes [51].

With the proposed interpretability framework, future research could analyze the video and audio modality to find shared patterns among samples that predicted the same label, to find out which features from the modalities give emotional cues, such as with the textual modality. Moreover, samples with the same highest contributing modality can be analyzed to find out which features globally are most influential within a modality. Moreover, patterns in misclassified samples for audio and video can be analyzed to identify potential weaknesses or biases in the model with respect to these modalities.

Furthermore, experiments can be conducted for the temporal dimension of the visual modality to highlight the frame with the highest contribution. The highest contributing frame can then be visualized to increase the interpretability of the spatial visualization.

Finally, the proposed interpretability method could be applied to other multimodal models for emotion recognition to analyze the robustness of these models and the framework itself.

5.5.4 Assessing the robustness of the interpretability framework

Although for evaluating their framework, in the original paper, MM-SHAP is compared to ablation methods and the Perceptual Degree, a quantitative comparison with the method using the attention mechanism is absent. Both methods, MM-SHAP and the attention mechanism, offer global and local insights, as can be read in Sections 2.5.1.1 and 2.5.1.3, and a comparative analysis considering both methods would increase the proposed method with the modified MM-SHAP. For each input token, Shapley values can be compared to attention values extracted from the last transformer block. Future research could study the extent of modality contribution at the sample, label, and dataset levels for SSE-FT on the MELD dataset according to the attention mechanism. This could determine whether there is a consistent agreement or divergence in results between the MM-SHAP framework and the attention mechanism when assessing modality contribution.

Chapter 6

Conclusion

In collaboration with the Dutch Institute of Sound & Vision, the current thesis explored the development of an interpretable multimodal for emotion recognition. The state-of-the-art multimodal model SSE-FT was trained and evaluated on the MELD dataset with its original architecture. Furthermore, an interpretability approach was proposed for analyzing multimodal models for emotion recognition and evaluating their robustness and limitations. This interpretability framework was implemented and evaluated on the model SSE-FT.

To achieve this, the MM-SHAP interpretability method was modified for the task of emotion recognition and extended to use text, video, and audio. Apart from evaluating multimodality, the proposed interpretability also provides detailed interpretability within the modalities by visualizing the Shapley values for each modality. Performance on the MELD test dataset was fairly low. The proposed interpretability framework found that SSE-FT relied solely on the textual backbone, as indicated by a T-SHAP score of 99%. Ablation studies verified these results and found that the textual backbone outperformed the model instance using all modalities. These results show that the model experiences 'uni-modal collapse' on the MELD dataset. The proposed interpretability framework demonstrated effectiveness in evaluating the multimodality of SSE-FT.

Moreover, the proposed approach describes a detailed method for analyzing the emotional cues captured by a model within each modality; however, due to the uni-modal collapse only the textual modality has been analyzed. Using the proposed interpretability framework to zoom in on the textual modality, it was found that each test sample had a high base value for the 'neutral' class. Due to the neutral bias of SSE-FT on the MELD dataset, most samples were predicted to be neutral. SSE-FT mostly relies on paralinguistic cues and exclamation words such as 'Oh' and 'God', to break the neutrality and assign an emotional label. In samples classified as 'neutral', it was found that words with high sentiment and their adjacent words contribute negatively to the 'neu-

tral' class, while these words contribute positively to emotional classes. This strongly suggests that the model has learned the semantics and context from the textual training data. Future work in Section 5.5.2 describes further experiments evaluating a model's ability to use context.

Research evaluating SSE-FT on the RAVDESS dataset results in good performance, even for the uni-modal models for audio and video [24]. This strongly suggests the uni-modal collapse is caused by the MELD dataset, namely, the model is not able to capture the more nuanced emotional cues from the audio and video modality and relies on the direct emotional cues from the textual modality. As mentioned before, the MELD dataset is imbalanced, as 48% of the training samples are neutral. The model is unable to discern between the underrepresented emotional classes. To use the audio and video modality, more clear samples for these classes are required, or the neutral bias in the training samples would have to be eliminated, as described in Section 5.5.1.

As an attempt to increase performance for the Sound & Vision archive, the SSL model RoBERTa was swapped for the Dutch variant RobBERT, which caused a decrease in performance. Training and evaluating RobBERT without the structure of SSE-FT resulted in the same accuracy. Fixing the imbalance in the textual data, eliminating 75% of 'neutral' samples improved performance. These results suggest that the SSL model RobBERT, although performing worse than RoBERTA, can be swapped as done in the current implementation; however, a well balanced dataset of good quality is still needed.

As it turns out, multimodal emotion recognition is quite difficult. The quality of the MELD dataset is too low for training an applicable multimodal model. Future work on improving the multimodal model for the Institute of Sound & Vision is described in Section 5.5.1. For a multimodal model to succeed in discerning emotions, it is most important to develop natural, balanced datasets that use clear rules on how to annotate emotions from multiple modalities. More specifically, a Dutch dataset for multimodal emotion recognition needs to be created, since currently there are none available. Such a dataset would significantly contribute to and inspire research in the field.

Bibliography

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in neural information processing systems*, 33:25–37, 2020.

[3] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[4] C.O. Alm, D. Roth, and Richard Sproat. Emotions from text: Machine learning for text-based emotion prediction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 347–354, 01 2005.

[5] Darari Nur Amali, Ali Ridho Barakbah, Adnan Rachmat Anom Besari, and Dias Agata. Semantic video recommendation system based on video viewers impression from emotion detection. In *2018 international electronics symposium on knowledge creation and intelligent computing (ies-kcic)*, pages 176–183. IEEE, 2018.

[6] Souha Ayadi and Zied Lachiri. A combined CNN-LSTM network for audio emotion recognition using speech and song attributs. In *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6. IEEE, 2022.

[7] Değer Ayata, Yusuf Yaslan, and Mustafa E Kamasak. Emotion recognition from multimodal physiological signals for emotion aware healthcare systems. *Journal of Medical and Biological Engineering*, 40:149–157, 2020.

[8] Kiavash Bahreini, Rob Nadolski, and Wim Westera. Towards multimodal emotion recognition in e-learning environments. *Interactive Learning Environments*, 24(3): 590–605, 2016.

[9] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[10] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[11] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Sequeira, Alexander Sutherland, and Stefan Wermter. The OMG-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1408–1414. IEEE, 2018. doi: 10.1109/IJCNN.2018.8489099.

[12] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *2003 Conference on computer vision and pattern recognition workshop*, volume 5, pages 53–53. IEEE, 2003.

[13] Ran Breuer and Ron Kimmel. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*, 2017.

[14] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[15] Sung-Woo Byun and Seok-Pil Lee. Human emotion recognition based on the weighted integration method using image sequences and acoustic features. *Multimedia Tools and Applications*, 80:1–15, 11 2021. doi: 10.1007/s11042-020-09842-1.

[16] Michele Cafagna, Lina M Rojas-Barahona, Kees Van Deemter, and Albert Gatt. Interpreting vision and language generative models with semantic visual priors. *Frontiers in Artificial Intelligence*, 6, 2023.

[17] Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and RBV Subramanyam. Benchmarking multimodal sentiment analysis. In *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18*, pages 166–179. Springer, 2018.

[18] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer, 2020.

[19] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[21] Hugo Carneiro, Cornelius Weber, and Stefan Wermter. Whose emotion matters? Speaking activity localisation without prior knowledge. *Neurocomputing*, 545:126271, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126271. URL https://www.sciencedirect.com/science/article/pii/S0925231223003946.

[22] Wallace Chafe. Punctuation and the prosody of written language. *Written communication*, 5(4):395–426, 1988.

[23] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Pier Luigi Mazzeo. ViTFER: Facial emotion recognition with vision transformers. *Applied System Innovation*, 5(4):80, 2022.

[24] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna, and Carlos M Travieso-González. Facial emotion recognition with inter-modality-attention-transformer-based self-supervised learning. *Electronics*, 12(2):288, 2023.

[25] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers, 2021.

[26] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan Su. This looks like that: Deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[27] Minping Chen and Xia Li. "SWAFN: Sentimental words aware fusion network for multimodal sentiment analysis". In *"Proceedings of the 28th International Conference on Computational Linguistics"*, pages 1067–1077, Barcelona, Spain (Online), December 2020. "International Committee on Computational Linguistics". doi: 10.18653/v1/2020.coling-main.93. URL https://aclanthology.org/2020.coling-main.93.

[28] Shizhe Chen and Qin Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 49–56, 2015.

[29] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

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

[31] Pieter Delobelle, Thomas Winters, and Bettina Berendt. Robbert: A dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*, 2020.

[32] James Deng and Clement Leung. Towards Learning a Joint Representation from Transformer in Multimodal Emotion Recognition. In *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings 14*, pages 179–188. Springer, 2021.

[33] Theo Deschamps-Berger, Lori Lamel, and Laurence Devillers. Investigating Transformer Encoders and Fusion Strategies for Speech Emotion Recognition in Emergency Call Center Conversations. In *Companion Publication of the 2022 International Conference on Multimodal Interaction*, pages 144–153, 2022.

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

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[36] Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. Unimodalities count as perspectives in multimodal emotion annotation. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*. CEUR Workshop Proceedings, 2023.

[37] Paul Ekman. Are there basic emotions? *Psychological review*, 99 3:550–3, 1992. URL https://api.semanticscholar.org/CorpusID:34722267.

[38] Paul Ekman and Wallace Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[39] Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[40] Muskan Garg, Seema Wazarkar, Muskaan Singh, and Ondřej Bojar. Multimodality for NLP-centered applications: Resources, advances and frontiers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6837–6847, 2022.

[41] Efthymios Georgiou, Charilaos Papaioannou, and Alexandros Potamianos. Deep hierarchical fusion with application in sentiment analysis. In *INTERSPEECH*, pages 1646–1650, 2019.

[42] Deepak Ghimire and Joonwhoan Lee. Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734, 2013.

[43] Maddalena Ghiotto. Archiving emotions: Conceptualizing multimodal emotion recognition in media culture. Master's thesis, UNIVERSITA' DI BOLOGNA, 2022-2023.

[44] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Contextual inter-modal attention for multimodal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3454–3466, 2018.

[45] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[46] Tamás Grósz, Dejan Porjazovski, Yaroslav Getman, Sudarsana Kadiri, and Mikko Kurimo. Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7026–7029, 2022.

[47] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[48] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[49] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 9:59800–59821, 2021.

[50] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[51] Aaishwarya Khalane, Rikesh Makwana, Talal Shaikh, and Abrar Ullah. Evaluating significant features in context-aware multimodal emotion recognition with xai methods. *Expert Systems*, page e13403, 2023.

[52] Dae Hoe Kim, Wissam J Baddar, Jinhyeok Jang, and Yong Man Ro. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Transactions on Affective Computing*, 10(2):223–236, 2017.

[53] Sangwon Kim, Jaeyeal Nam, and Byoung Chul Ko. ViT-NeT: Interpretable vision transformers with neural tree decoder. In *International Conference on Machine Learning*, pages 11162–11172. PMLR, 2022.

[54] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021.

[55] Joseph Kruskal. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, 25(2):201–237, 1983.

[56] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.

[57] Harold William Kuhn and Albert William Tucker. *Contributions to the Theory of Games*. Number 28. Princeton University Press, 1953.

[58] Sanghyun Lee, David K. Han, and Hanseok Ko. Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification. *IEEE Access*, 9:94557–94572, 2021. doi: 10.1109/ACCESS.2021.3092735.

[59] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021.

[60] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[61] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable mltimodal emotion reasoning. *arXiv preprint arXiv:2306.15401*, 2023.

[62] Hugo Liu, Henry Lieberman, and Ted Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 125–132, 2003.

[63] Pengfei Liu, Kun Li, and Helen Meng. Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition. *arXiv preprint arXiv:2201.06309*, 2022.

[64] Cristina Luna-Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M. Montero, and Fernando Fernández-Martínez. Multimodal emotion recognition on RAVDESS dataset using transfer learning. *Sensors*, 21(22), 2021. ISSN 1424-8220. doi: 10.3390/s21227665. URL https://www.mdpi.com/1424-8220/21/22/7665.

[65] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[66] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467, 2022.

[67] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021.

[68] Pranava Madhyastha, Josiah Wang, and Lucia Specia. Defoiling foiled image captions. *arXiv preprint arXiv:1805.06549*, 2018.

[69] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[70] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012. doi: 10.1109/T-AFFC.2011.20.

[71] Tao Meng, Yuntao Shou, Wei Ai, Nan Yin, and Keqin Li. Deep imbalanced learning for multimodal emotion recognition in conversations. *arXiv preprint arXiv:2312.06337*, 2023.

[72] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[73] A. Milton, Sharmy Roy, and S. Tamil Selvi. SVM scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, 69(9), 2013.

[74] Emily Ohman, Kaisla Kajava, Piao Hui, and Jörg Tiedemann. Emotion preservation in translation: Evaluating datasets for annotation projection. 10 2020.

[75] Letitia Parcalabescu and Anette Frank. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks, 2023.

[76] Nivedita Patel, Shireen Patel, and Sapan H Mankad. Impact of autoencoder based compact representation on emotion detection from audio. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–19, 2022.

[77] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *"Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)"*, pages 1532–1543, "Doha, Qatar", October 2014. "Association for Computational Linguistics". doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[78] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[79] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[80] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.

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

[82] Mohammad Rabiei and Alessandro Gasparetto. A methodology for recognition of emotions based on speech analysis, for applications to human-robot interaction. an exploratory study. *Paladyn, Journal of Behavioral Robotics*, 5(1): 000010247820140001, 2014.

[83] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[84] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 1135–1144, 2016.

[85] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[86] Gaurav Sahu. Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022*, 2019.

[87] Ruslan Salakhutdinov and Geoffrey Hinton. Deep Boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455. PMLR, 2009.

[88] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[89] Saranya and Subhashini. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7:100230, 2023. ISSN 2772-6622. doi: https://doi.org/10.1016/j.dajour.2023.100230. URL https://www.sciencedirect.com/science/article/pii/S277266222300070X.

[90] Yutaka Sasaki et al. The truth of the F-measure. *Teach tutor mater*, 1(5):1–5, 2007.

[91] Nobuo Sato and Yasunari Obuchi. Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2(3):835–848, 2007.

[92] Anvita Saxena, Ashish Khanna, and Deepak Gupta. Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, 2(1):53–79, 2020.

[93] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[94] Lloyd S Shapley et al. A value for n-person games. 1953.

[95] Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. MEmoR: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 493–502. ACM, 2020.

[96] Amir Shirian and Tanaya Guha. Compact graph architecture for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6284–6288. IEEE, 2021.

[97] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285, 2020. doi: 10.1109/ACCESS.2020.3026823.

[98] Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28, 2015.

[99] Lang Su, Chuqing Hu, Guofa Li, and Dongpu Cao. MSAF: Multimodal split attention fusion, 2020.

[100] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[101] Licai Sun, Mingyu Xu, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng. Multimodal emotion recognition and sentiment analysis via attention enhanced recurrent model. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, MuSe '21, page 15–20, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386784. doi: 10.1145/3475957.3484456. URL https://doi.org/10.1145/3475957.3484456.

[102] Jacopo Teneggi, Alexandre Luster, and Jeremias Sulam. Fast hierarchical games for image explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4494–4503, 2022.

[103] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.

[104] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656.

[105] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, page 1823. NIH Public Access, 2020.

[106] Christiana Tsiourti, Astrid Weiss, Katarzyna Wac, and Markus Vincze. Multimodal integration of emotional signals from voice, body, and context: Effects of (in)congruence on emotion recognition and attitudes towards robots. *International Journal of Social Robotics*, 11, 08 2019. doi: 10.1007/s12369-019-00524-z.

[107] Patti M Valkenburg, Jochen Peter, and Joseph B Walther. Media effects: Theory and research. *Annual review of psychology*, 67:315–338, 2016.

[108] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[109] Sutong Wang, Yunqiang Yin, Dujuan Wang, Yanzhang Wang, and Yaochu Jin. Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. *IEEE transactions on cybernetics*, 52(12):12623–12637, 2021.

[110] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223, 2019.

[111] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*, 2021.

[112] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[113] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation, 2019.

[114] Chen Yu and Dana Ballard. On the integration of grounding language and learning objects. In *AAAI*, volume 4, pages 488–493. Citeseer, 2004.

[115] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha

Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115.

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

[117] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind Picard. Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–30, 2020.

[118] Guihua Zhang, Jian-Wei Lin, Ji Wang, Jie Ji, Ling-Ping Cen, Weiqi Chen, Peiwen Xie, Yi Zheng, Yongqun Xiong, Hanfu Wu, Dongjie Li, Tsz Ng, Chi Pang, and Mingzhi Zhang. Automated multidimensional deep learning platform for referable diabetic retinopathy detection: A multicentre, retrospective study. *BMJ open*, 12: e060155, 07 2022. doi: 10.1136/bmjopen-2021-060155.

[119] Yazhou Zhang, Qiuchi Li, Dawei Song, Peng Zhang, and Panpan Wang. Quantum-inspired interactive networks for conversational sentiment analysis. 2019.

[120] Yifei Zhang, Neng Gao, and Cunqing Ma. Learning to Select Prototypical Parts for Interpretable Sequential Data Modeling, 2023.

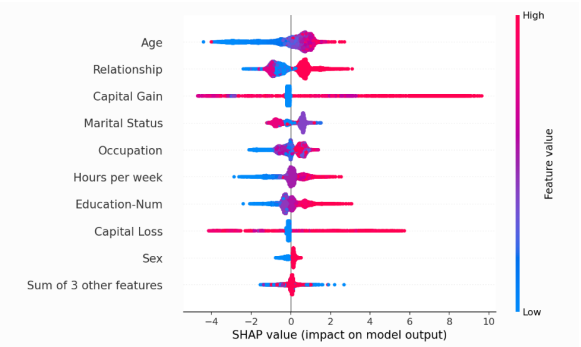
[121] Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. MEmoBERT: Pre-training model with prompt-based learning for multimodal emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4703–4707. IEEE, 2022.

[122] Pablo Zinemanas, Martín Rocamora, Marius Miron, Frederic Font, and Xavier Serra. An interpretable deep learning model for automatic sound classification. *Electronics*, 10(7), 2021. ISSN 2079-9292. URL https://www.mdpi.com/2079-9292/10/7/850.

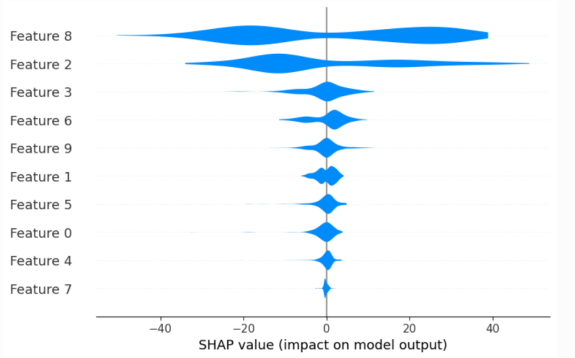
Appendix A

SHAP Visualization Plots

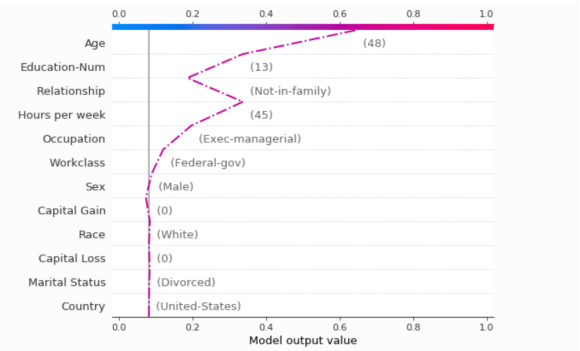
(1) Bar Plot



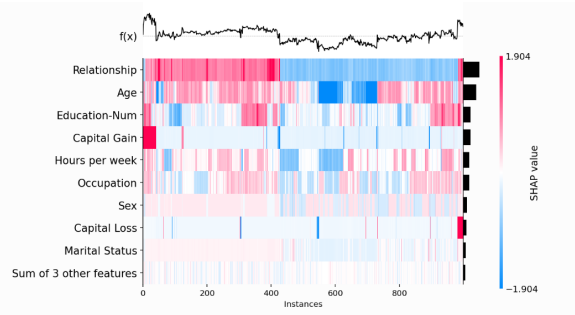
(2) Beeswarm Plot



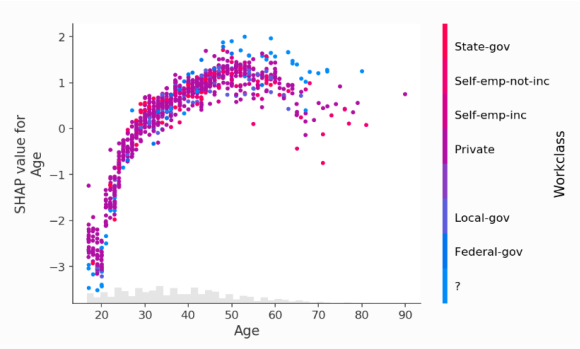
(3) Violin Plot



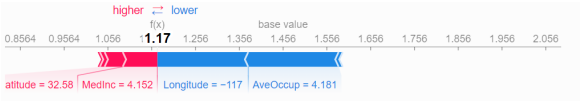
(4) Decision Plot



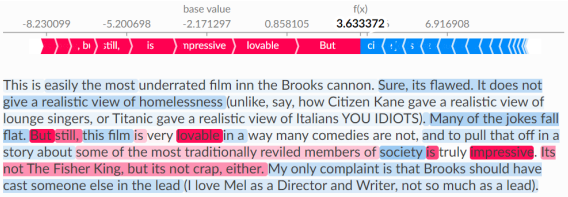
(5) Heatmap Plot



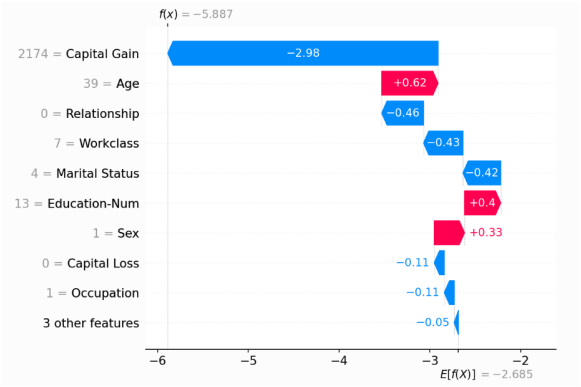
(6) Dependence Scatter Plot



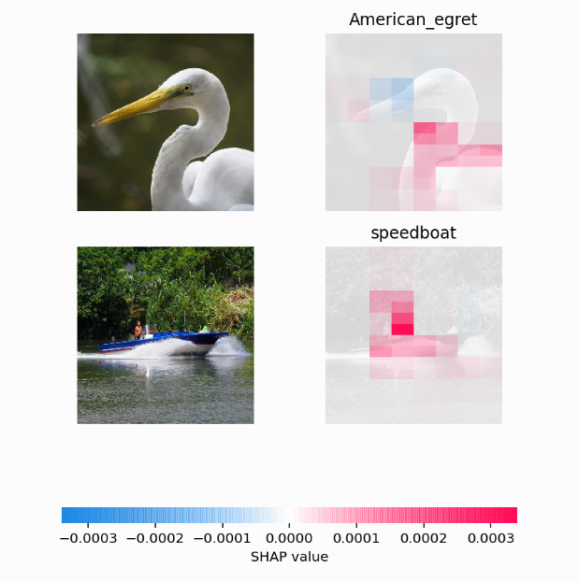
(7) Force Plot



(8) Text Plot



(9) Waterfall Plot



(10) Image Plot