REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

MULTIMODAL EMOTION TRACKING FROM VIDEOS USING IMAGE, AUDIO, AND TEXT

19011010 — Beyda GÜLER
20011055 — Şeymanur KORKMAZ

SENIOR PROJECT

Advisor
Assoc. Dr. Ali Can KARACA

June, 2024

ACKNOWLEDGEMENTS

We would like to sincerely thank our esteemed advisor for our project, Assoc. Prof. Dr. Ali Can Karaca. He has been instrumental in helping us along the way, giving us invaluable guidance at crucial times, and illuminating our path with his knowledge and experience as we shape our careers.

We also thank Res. Ass. Kübra Adalı for her help and patience in fixing our challenges during the course of our studies.

Beyda GÜLER
Şeymanur KORKMAZ

TABLE OF CONTENTS

LIST OF SYMBOLS

%	Percentage Sign
$	Dollar Sign

LIST OF ABBREVIATIONS

avg	Average
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLSTM	Bidirectional Long Short-Term Memory
BoW	Bag of Words
GPU	Graphical User Interface
IEMOCAP	The Interactive Emotional Dyadic Motion Capture
LSTM	Long Short-Term Memory
MFCC	Mel-frequency CepstruM Coefficient
MLM	Masked Language Modeling
NER	Named Entity Recognition
NLP	Natural Language Processing
NSP	Next Sentence Prediction
OpenSMILE	Open Speech and Music Interpretation by Large Space Extraction
PCA	Principal Componenyt Analysis
RAM	Random Access Memory
RNN	Recurrent Neural Networks
SemEval	Semantic Evaluation
SVM	Support Vector Machines
TF-IDF	Term Frequency — Inverse Document Frequency
TPU	Tensor Processing Units
URL	Uniform Resource Locator
YAMNet	Yet Another MusicNet

Wav2Vec	Wave to Vector
3D CNN	3 Dimensional Convolutional Network

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

MULTIMODAL EMOTION TRACKING FROM VIDEOS USING IMAGE, AUDIO, and TEXT

Beyda GÜLER
Şeymanur KORKMAZ

Department of Computer Engineering
Senior Project

Advisor: Assoc. Dr. Ali Can KARACA

Each type of data provides us with different types of information. For example, text-based data provides information in terms of grammar and context, audio-based data includes information such as intonation, rhythm and rate of speech, and video data includes different types of information such as gestures, visual cues and facial expressions. When these features are combined, the information provided by each type is integrated, resulting in richer new information. Hence the need for multi-modal emotion tracking, which is the starting point of this study.

In this study, multi-modal emotion detection and tracking is performed on two different datasets. One of the datasets is highly unbalanced and the other is relatively more balanced to see how the study would work under different conditions.

Text, audio and video models are created on the datasets and outputs are obtained, and then this information is combined to create new outputs from multi-modal models. As a result, it is seen how multi-modal working contributes to knowledge acquisition.

Keywords: emotion detection, emotion tracking, multi-modal, data, model, knowledge acquisition

ÖZET

PROJE KİTABI HAZIRLAMA ESASLARI VE LATEX TASLAĞI KULLANIMI

Beyda GÜLER
Şeymanur KORKMAZ

Bilgisayar Mühendisliği Bölümü
Bitirme Projesi

Danışman: Doç. Dr. Ögr. Üyesi Ali Can KARACA

Verinin her çeşidi bizlere farklı türde bilgiler sunmaktadır. Örneğin; metin tabanlı veriler dil bilgisi ve bağlam açısından bilgi edinimi sağlarken ses tabanlı veriler tonlama, ritim ve konuşma hızı gibi bilgileri, video içeren veriler ise jest, görsel ipuçları ve mimikler gibi farklı türde bilgileri içerir. Bu özellikler birleştirildiğinde ise her türün sunduğu bilgiler entegre edilerek daha zengin yeni bir bilgi elde edilmiş olur. Bu çalışmanın doğuş noktası olan çok modlu duygu takibine bu sebeple ihtiyaç duyulmuştur.

Bu çalışmada iki farklı veri seti üzerinde çok modlu duygu tespiti ve takibi yapılmaktadır. Kullanılan veri setlerinden biri oldukça dengesiz ve diğeri de nispeten daha dengeli seçilerek farklı koşullarda çalışmanın nasıl neticeleneceği incelenmektedir.

Veri setleri üzerinde metin, ses ve video ayrı olacak şekilde modeller oluşturularak çıktılar elde edilmekte ve daha sonra bu bilgiler birleştirilerek çok modlu modellerden yeni çıktılar oluşturulmaktadır. Bunun sonucunda çok modlu çalışmanın bilgi edinimine ne yönde katkı sağladığı görülmektedir.

Anahtar Kelimeler: duygu tespiti, duygu takibi, çok modlu, veri, model, bilgi edinimi

1
Introduction

This section provides an overview of the topics covered in the project.

1.1 Natural Language Processing

Through the use of machine learning techniques, computers can now interpret, process, and comprehend human language thanks to natural language processing, or NLP. Organizations nowadays are dealing with massive amounts of text and voice data from a variety of communication channels, including emails, texts, social media news feeds, videos, and audio files. They automatically process this data, interpret the message's intent or emotion, and react to human communication in real time using natural language processing (NLP) software.

Natural languages and machine languages—also known as programming languages—are the two categories of languages used in the computer industry. Analysis, comprehension, and/or replication of natural languages' regular structure are the goals of natural language processing. NLP blends computer linguistics, deep learning, machine learning, statistics, and rule-based modeling of human language.

In addition to reading human language, machines that use natural language processing (NLP) can also understand and interpret it. NLP allows machines to perform tasks like sentiment analysis, speech recognition, and automatic text summarization. It also allows machines to comprehend language, both spoken and written.

Many people engage with these systems on a daily basis, whether they realize it or not. Natural language processing technology, for instance, powers chatbots, virtual personal assistants, Apple Siri, and Google Assistant. The advantages of natural language processing extend beyond the personal conveniences it affords us on a daily basis. Additionally, it has expedited its marketing research projects and made significant contributions to the private sector's ground-breaking understanding of consumer needs. [1].

Beyond our interactions with virtual assistants like Alexa or Siri, natural language processing, or NLP, has become an indispensable part of our everyday lives. Among the notable uses is Bayesian spam filtering, which finds commonalities in spam emails' subject lines by applying statistical natural language processing techniques. Speech-to-text conversion, an NLP-based feature, also makes it simple to read automatic transcriptions of voicemails in email inboxes or smartphone apps. Furthermore, users are essentially using natural language processing (NLP) techniques for tasks like search, topic modeling, entity extraction, and content categorization when they use integrated search bars on websites or select suggested tags for topics, entities, or categories. These real-world examples demonstrate how NLP has a broad impact on enhancing many facets of our digital interactions.

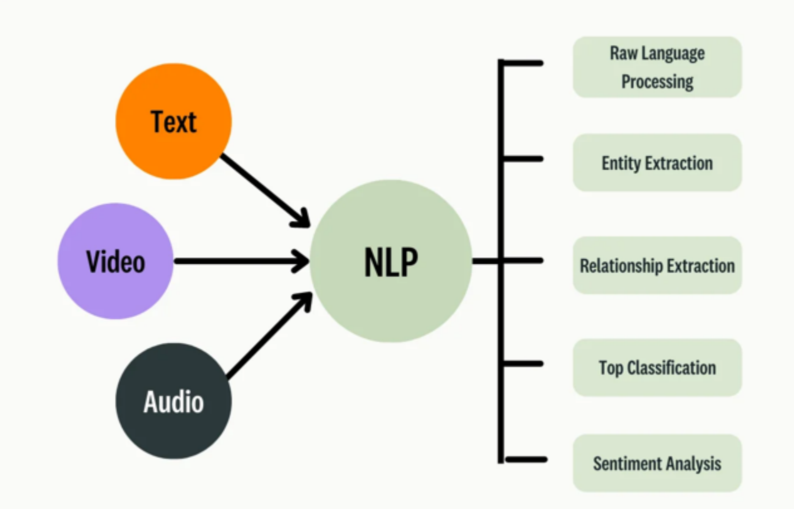


Figure 1.1 Natural Language Processing [2]

The following categories describe the stages of natural language processing: morphology, pragmatic analysis (pragmatics), semantic analysis (semantics), lexical analysis (syntax), and discourse analysis (discourse). Semantic analysis is used in our project, "Multimodal Emotion Tracking from Videos Using Image, Audio, and Text," to extract emotion from text.

1.2 Text Processing

The automated process of evaluating and classifying unstructured text data to extract insightful information is known as text processing. Text processing tools make use of artificial intelligence subfields such as machine learning and natural language processing (NLP) to automatically comprehend human language and extract value from text data. Natural language processing is the ability of a computer to comprehend human language in a useful way, whereas text processing only refers to the analysis, creation, and manipulation of text.

The following explains how NLP and text processing are related: In essence, text processing is followed by natural language processing.For instance, a machine learning model could be trained in advance to find instances of positive or negative sentiment words in order to perform a basic sentiment analysis. Since the model is only searching for words that it has been programmed to look for and isn't understanding the words, this would be considered text processing. Translating complete sentences into another language would be an example of a natural language processing model. The computer must comprehend the meaning of the sentences in order to translate them accurately because syntax differs between languages. However, even though NLP has advanced beyond text processing, it always has text processing involved as a step in the process [3].

What applications does text processing have?

- **Topic Analysis:** Using this method, extensive text collections are interpreted and categorized into topics or themes.
- **Sentiment Analysis:** This feature categorizes text's emotional undertones as positive, negative, or neutral by automatically detecting them.
- **Intent Detection:** The intent, purpose, or objective of the text is identified by this classification model. It might ascertain, for instance, if the goal is to obtain information, make a purchase, or cancel a subscription to the business.
- **Language Classification:** Text is categorized according to the language in which it was written.

1.3 Image Processing

The process of converting an image into a digital format and carrying out specific operations in order to extract some valuable information is known as image processing. Generally speaking, the image processing system applies specific preset signal processing techniques to all images as 2D signals. [4].

The use of computer algorithms to modify digital images is the focus of the "digital image processing" technique class. It is an essential preprocessing step for many applications, such as object detection, face recognition, and image compression.

Image processing is the process of improving an already-existing image or extracting significant information from it. This is crucial for a number of Deep Learning-based Computer Vision applications, as these kinds of preprocessing can significantly increase a model's performance. Another use, particularly in the entertainment sector, is image manipulation, such as the addition or removal of objects. [5].

These days, companies in many different industries use the rapidly developing field of image processing systems. This technology stands for an essential field of study in computer science and engineering.

There are five categories for the purpose of image processing. These are the following:

- **Visualization** – Seeing objects that are difficult to see;
- **Image sharpening and restoration** – Increasing the quality of noisy images
- **Image acquisition** – looking for visually appealing and high-quality photos
- **Pattern Recognition** – different objects in an image
- **Image Recognition** – differentiating objects in an image [6]

Image processing has numerous applications. Specifically, it can be applied to enhance image quality, increase readability, and enable computer comprehension.

An essential component of artificial intelligence (AI) is image processing. Artificial intelligence (AI) systems can learn to recognize objects and patterns by comprehending and processing images. They are able to make wiser choices and complete tasks more successfully as a result.

Compressing images to reduce their size on storage devices or during network transmission is another use for image processing. This is crucial for cutting expenses and guaranteeing speedy and effective data transfer.

'Multimodal Emotion Tracking from Videos Using Image, Audio, and Text' is our project that uses images to extract emotions. By examining facial expressions, the subfield of image processing technology known as emotion recognition seeks to identify the emotional states of people. This technology typically analyzes digital photos or photographs to assess visual characteristics like eyebrow position, eye health, and facial expressions. Advanced algorithms can determine an individual's emotional states, such as happiness, sadness, fear, or surprise, by examining these traits. Applications for emotion recognition could be found in a wide range of fields, including marketing, security, and human-machine interactions.

1.4 Audio Processing

Audio processing is a multifaceted domain that operates on the frontier of technological advancements to capture, manipulate, analyse, and synthesise sound. Deep tech solutions allow us to unlock capabilities beyond traditional sound processing techniques [7].

Innovations in emerging technologies include applications such as real-time speech recognition and translation, complex voice synthesis, and even emotion and context recognition. Here are some use cases of new technologies and their applications in audio:

- **Watermarking:** Methodology for invisibly watermarking audio or video streams with a unique receiver identifier. Ideally, the watermark should be insensitive to compression and transformations of the audio or video. Used to track the potential source of data leaks by retrieving the receiver ID from the leaked stream.
- **Speech Denoising:** Particularly useful in noisy environments where multiple background sounds are captured. Helps separate the speech audio signal from background noise, improving clarity and intelligibility of transmitted speech.
- **Text-to-Speech and Voice Cloning:** Speech synthesis combined with voice cloning allows for synthesizing a specific person's voice. Useful when the original speaker is unavailable or costly. Realistic, expressive speech synthesis is crucial for a human-like experience for the listeners.
- **Video Analysis:** For sports events, analysis of videos from multiple cameras to reconstruct 3D details of the game (tennis, football, basketball). Involves tracking players among numerous cameras, rebuilding their body movement, and summarizing the most important statistics [7].

Technological developments in deep learning, machine learning, and neural networks are going to improve our ability to analyze and interpret complex audio. This development creates opportunities for large-scale real-time processing, speech recognition, noise cancellation, and synthesis. Along with advancements in Automated Speech Recognition (ASR) and Natural Language Processing (NLP), the continuous collection of large datasets holds promise for further improving the precision and efficacy of audio processing systems.

The importance of audio processing is emphasized by strong trends and facts. With a projected value of USD 10.42 billion in 2022 and a strong Compound Annual Growth

Rate (CAGR) of 24.8%, the global speech and voice recognition market is expected to grow to USD 59.62 billion by 2030. According to a 2021 report, web conference transcription—which is augmented by speech and voice recognition technology—has a significant 44% market share in the voice technology space. Although there have been significant developments in Audio-Visual Speech Recognition (AVSR) techniques in the past ten years, audio-visual speech decoding problems still exist. An automatic music signal mixing system based on one-dimensional Wave-U-Net autoencoders is highlighted in a 2023 research publication, which further highlights the significant impact of deep learning integration on music signal processing. All of these facts highlight how important audio processing has been in forming and developing different technological fields. [7].

1.5 The Goal of the Project

Recognizing emotions directs one's reaction and behavior toward potentially amicable or dangerous individuals. Additionally, effective interpersonal communication depends on the ability to recognize emotions in others. We process both static and dynamic information, such as body language and facial expressions, to detect emotions [8].

AI that can recognize and analyze various facial expressions in order to apply them to extra information that is supplied to them is known as emotion recognition. This helps investigators and interviewers alike, as it makes it possible for authorities to utilize technology alone to determine an individual's feelings.

The system can be fed with a variety of inputs for analysis. The main areas of emotion recognition are as follows:

Video: It is a significant dataset in the field of affective computing. To learn how to employ them in emotion recognition, a great deal of study is still being done. Mobile phone cameras are one example of contemporary software.

Image: It is one of the key sets of available sets in emotion identification, just like video. According to some study, automatically classifying photographs based on their emotions and categorizing video segments into different genres can both benefit from the use of classified emotions in photos.

Speech: Speeches are typically written down and analyzed as texts. It would not be possible to recognize emotions using this method. Numerous studies are continuously being conducted to investigate the use of speech rather than written texts for applications that identify emotions.

Text: The most common data format for texts is the Document-Term Matrix, or DTM. As it use single words, it is inappropriate for identifying emotions. The way a text is seen depends on its tone, punctuation, and other elements, all of which are taken into account by academics as they continue to explore novel text data structures.

Conversation: Acquiring emotions from conversations involving two or more people is the main goal of emotion recognition in conversation. Typically, the datasets under this category come from social media platforms' free samples. Nonetheless, other obstacles persist in this domain, such as the occurrence of irony in a discussion, the alteration of the interlocutor's emotions, and conversational-context modeling.

Giving only one of these input types to the system may not be sufficient for emotion detection. Therefore, working in a combined manner on more than one input type will be more useful for emotion detection. Our project 'Multimodal Emotion Tracking from Videos Using Image, Audio and Text' aims to predict emotions in a dialogue series and find sentences that create emotional changes through data in different modalities such as video, audio and text.

1.6 Dataset

We perform experiments on two benchmark: SemEval [9] and IEMOCAP [10].

SemEval (Semantic Evaluation) 2024 Task 3 Dataset: SemEval dataset, contains the video clips from the Friends sequence and the training data files, where all the examples are stacked in a list and each one is stored in a dictionary. The dataset contains a total of 13.587 videos and text files containing the dialogues in these videos. The sounds of the given videos were separated and an audio folder was created for audio processing. The dataset is divided into 9782 for train, 1359 for validation and 2446 for test. The dataset consists of 7 classes representing 7 different emotions. These are joy, surprise, anger, sadness, sadness, disgust, fear. There is also an unbalanced distribution of the data set. As can be seen in Figure 1.2, the joy class is superior to other classes. This is one of the factors that can negatively affect the success of the model.

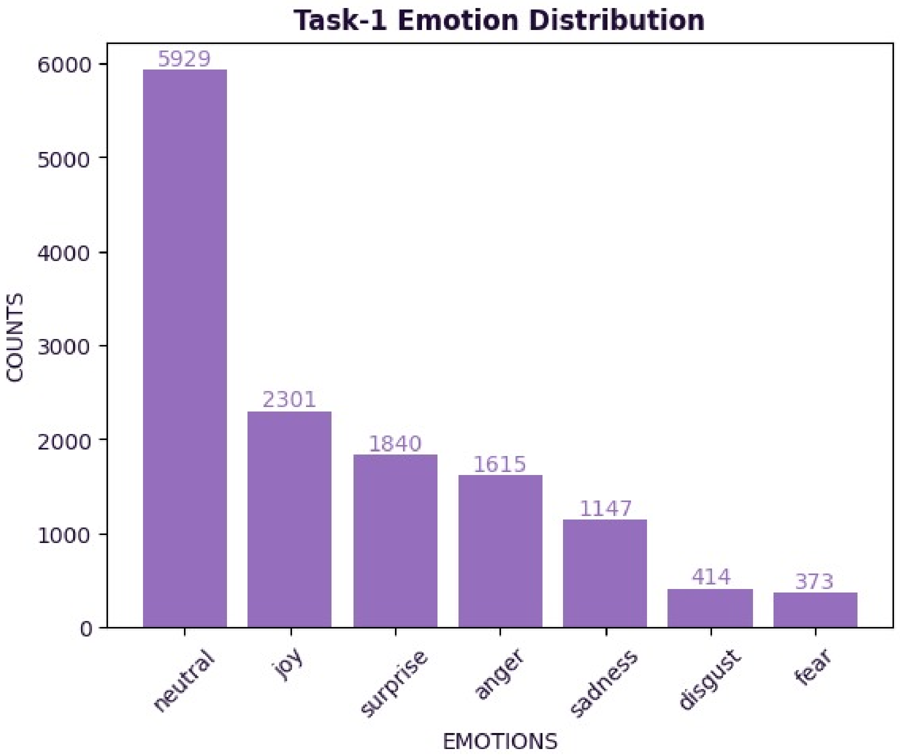


Figure 1.2 SemEval Dataset Class Distribution

IEMOCAP (The Interactive Emotional Dyadic Motion Capture) Database: Videos of dyadic conversations between pairs of ten speakers make up the IEMOCAP database. Videos of dyadic conversations between pairs of ten speakers make up the IEMOCAP database. Each pair is given a variety of scenarios to discuss after being divided into five sessions. Videos are divided into speech segments and annotated with precise emotion classifications. For the classification task, we take into account six such categories: neutral, excitement, frustration, sadness, happiness, and anger. The first eight speakers from sessions 1-4 are used to select the training set, and session 5 is reserved for testing.Each pair is given a variety of scenarios to discuss after being divided into five sessions. Videos are divided into speech segments and annotated with precise emotion classifications. For the classification task, we take into account six such categories: neutral, excitement, frustration, sadness, happiness, and anger. The data set's class distribution is depicted in Figure 1.3. 5202 for train, 744 for validation, and 1487 for test make up the dataset.

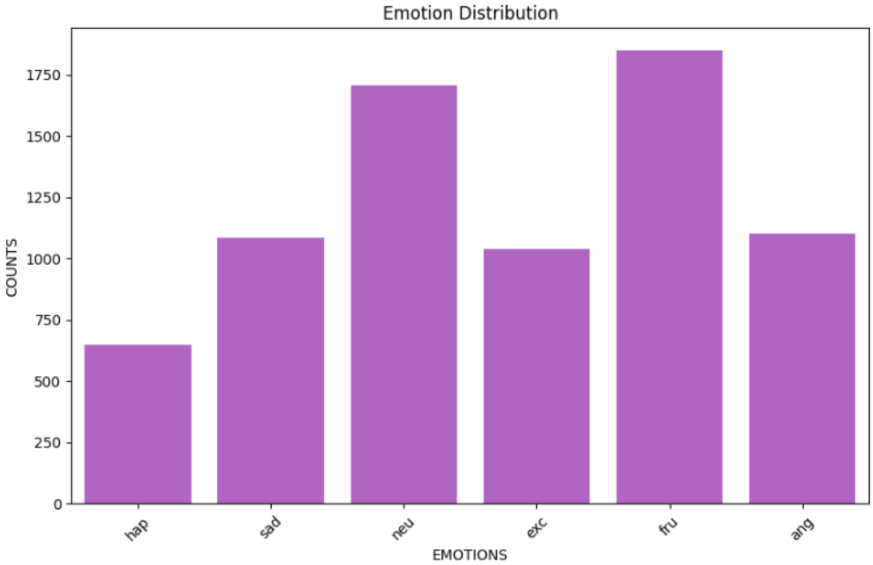


Figure 1.3 IEMOCAP Dataset Class Distribution

2
Preliminary Review

In this section, a preliminary review of the project is made and decisions that will guide the course of the project are discussed.

2.1 Need For the Project

Sentiment analysis is an important topic in order to understand the data correctly and give the right response. To give an example from real life, if we misunderstand the message conveyed to us, we inhibit communication and give wrong answers to the other person. In this context, many technologies are being developed today to meet the communication need more effectively (such as voice and video calling). The situation is similar when we consider this problem at the level of computers. Correct analysis of the transmitted data is vital for further processes. Therefore, using only text or only voice may be insufficient when processing real life data. At this point, the need for this project emerges. Emotion analysis can be done by text processing by transcribing a speech, but in this case, we do not take into account important parameters such as gestures, facial expressions, tone of voice, etc. In order to perform emotion analysis in the most accurate way, it is required to combine verbal, written and visual inputs in most effective way.

2.2 Project Scope

The expected achievements of the system within the scope of our project are as follows:

- Downloading, analyzing, and reporting the characteristics of the dataset.
- Extraction of text features.
- Extraction of audio features.
- Extraction of video features.

- Emotion detection using only text.
- Emotion detection using only audio.
- Emotion detection using only video.
- Combining features
- Emotion classification by combined features.
- Performance measurement.

3
Feasibility

This section examines the paths that can be followed in the development of the application. Details such as the platform on which the application will be developed, its legal and economic feasibility will be understood.

3.1 Technical Feasibility

Software and hardware feasibility will be described under this section.

3.1.1 Software Feasibility

Python programming language version 3.10.12 was used in the development of the project. Google Colab, which was found to be the most suitable development environment, was used and the codes were compiled in this environment. Python libraries used in the development of artificial intelligence models are also completely free and open source. The use of open source libraries had a positive impact on the implementation of the system.

3.1.2 Hardware Feasibility

The project is implemented on Windows 10 operating system. Since the development of the project is based on Google Colab, the CPU and GPU of the computer as well as the TPU and GPU support provided by Colab are used. Colab is more advantageous in this respect compared to Jupyter Notebook. This is an important point since we are dealing with large-scale data and it helps to shorten the working time. Also, Google Colab also offers support for RAM and storage space.

3.2 Legal Feasibility

The SemEval data set used was taken from the Google Drive [11] shared by the Semeval 2024 Task-3 (Multimodal Emotion Cause Analysis in Conversations) competition owners and the IEMOCAP data set was also taken from Google Drive [12]. Since both are open source, there are no legal problems in use. The project does not keep any data that would violate laws and regulations in any way. All libraries and frameworks used are open source and free.

3.3 Labor & Time Feasibility

It was carried out by 2 people in 3 and a half months. In Figure 3.1, the time required and spent to complete the tasks is shown.

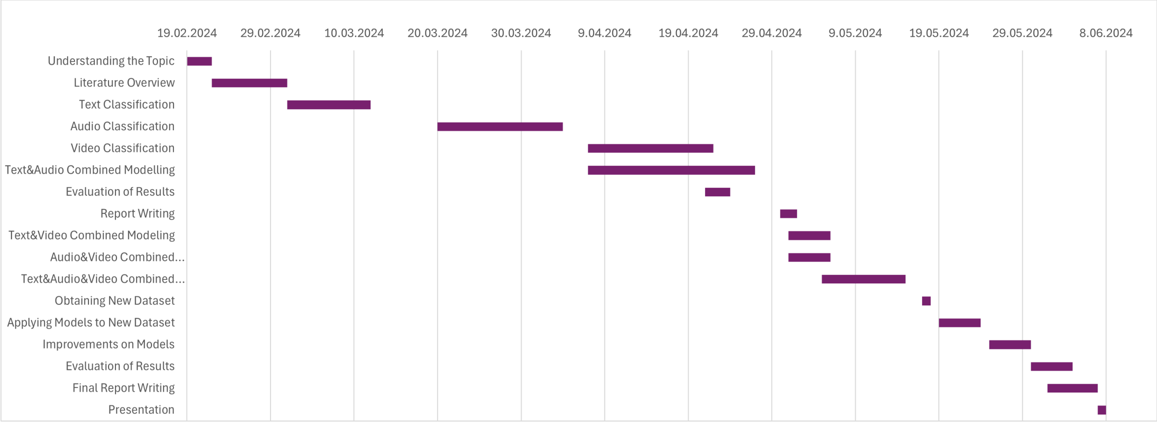


Figure 3.1 Labor & Time Feasibility

3.4 Economical Feasibility

The PRO version of Google Colab is used as the development environment. The cost of this version is 9.99$. Deep learning and machine learning methods used as software tools are also free of charge. The dataset is open source and free of charge in SemeEval. HP Pavilion 15 and Lenovo IdeaPad L340 personal computers were used as hardware. These computers meet the system requirements. The detailed cost table is as shown in the Table 3.1.

Table 3.1 Economical Feasibility Table

Tool	Quantity	Price	Cost
Lenovo IdeaPad L340	1	1500$	1500$
HP Pavilion 15	1	1100$	1100$
Google Colab PRO	2	9.99$	19.98$
Total Cost			2619.98$

4
System Analysis

This section describes the system requirements, goals and performance metrics.

4.1 Requirements

In this project, the need for labeled data was met with the data set provided by Semeval and IEMOCAP. In addition, tools and libraries (such as TensorFlow, NumPy, Pandas etc.) are needed to model this data with various machine learning and deep learning methods.

In order to import and use libraries, development environments such as Jupyter Notebook, Anaconda Navigator or Google Colab are required. Among these environments Google Colab is preferred due to its higher capacity. GPUs and TPUs provided by Google Colab are used as accelerators. Python coding language is used.

4.2 Goals

The main goal of our study is to perform sentiment analysis on unclassified inputs. In this regard, firstly the audio, text and visual inputs we have will be given to the model separately and then various combinations of inputs will be given to the models and the most optimal model will be used to perform sentiment analysis. Thus, it will be determined which factor is more distinctive on emotion detection. Also, data will follow the necessary pre-processing steps before modelling.

4.3 Performance Metrics

To determine the system success, predicted labels from several deep learning and machine learning algorithms are compared to the test dataset. Success is measured using metrics including accuracy, F-1 weighted average, and F-1 macro average. For this project, the F-1 macro average is primarily used to calculate success.

The rationale behind taking into account the macro average is that precision can lead to misleading findings in cases where the class distribution is not balanced. The weighted average, on the other hand, increases the significance of large class performance. In this study, the macro average is used since it is desirable to account for the performance imbalance across classes and not overlook the performance of minor classes.

5
System Design

This section describes the software design.

5.1 Software Design

The software design stages are as in Figure 5.1.

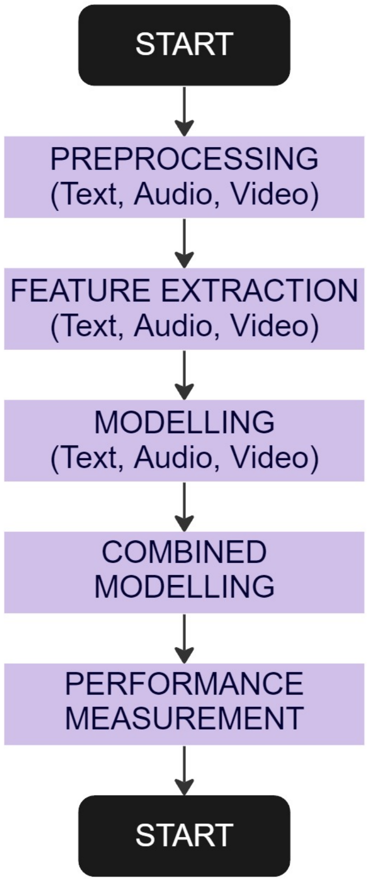


Figure 5.1 Software Design Flowchart [13]

5.1.1 Preprocessing

- **Text Preprocessing** Natural language text data is typically noisy and unstructured. Therefore, text preprocessing is essential to converting unstructured, messy text data into a format that can be used to train machine learning models efficiently, producing better outcomes and insights. [14]. The basic preprocessing steps applied to the texts are as follows:
  - Lowercasing
  - Eliminating special characters and punctuation
  - Removing numbers
  - URL removal
  - Cleaning spaces
- **Video Preprocessing**
  - Extracting frames from each video
  - Padding and resizing frames from a video

5.1.2 Feature Extraction

The technique of extracting pertinent information or features from unprocessed data to provide a more condensed representation while preserving the most crucial aspects of the original data is known as feature extraction. Feature extraction is an essential tool in many domains, such as computer vision, signal processing, machine learning, and natural language processing, as it helps to deconstruct complex data and prepare it for analysis or model training. Text, audio, and video feature extraction were all done as part of our project. Using the characteristics that were retrieved for text, audio, and video in classification models, classification is carried out.

5.1.2.1 Text Feature Extraction

Text feature extraction, which is of great importance for NLP applications, involves the steps of analysing and processing text data. There are basic methods such as BoW and TF-IDF that provide quick and easy solutions, as well as word embedding and transducer-based models that provide more precise and contextual representations.

Some text feature extraction methods are as follows.

1. **BERT:** Natural language processing (NLP) was transformed by Google's BERT (Bidirectional Encoder Representations from Transformers) language paradigm.

Introduced in 2018, BERT extracts context information from both sides of the text to better extract word meaning due to its bidirectional approach. This is an important distinction that distinguishes BERT from previous unidirectional models. Evaluating the meaning of a word in relation to all the words around it enables a more accurate and richer contextual representation of meaning to be produced.

BERT is trained on two primary tasks:

- **Masked Language Model (MLM):** During training, some words are replaced with a masking symbol ([MASK]) and the model is asked to predict these masked words. This allows the model to learn bidirectional context.
- **Sentence Pairs Classification (Next Sentence Prediction - NSP):** The model predicts whether two sentences are consecutive or not. This task helps learn relationships between sentences.

BERT produces contextual embeddings for words and sentences when extracting text features. These embeddings represent the meanings of words and sentences and through these representations the following steps can be performed.

- **Text Processing:** Tokenization is the process of breaking up input texts into words and subwords respectively. By dissecting words into their constituent pieces, BERT's unique tokenizer pulls more precise contextual information.
- **Input Representation:** BERT takes word, position and segmentation features as input. Through these representations, the model fully captures each word.
- **Feature Extraction:** Contextual representations are high-dimensional vectors that are created from the input as it moves through the layers of BERT. The text's syntactic and semantic elements are carried by these vectors.
- **Usage:** Contextual embeddings that are produced can be applied to a variety of natural language processing tasks, including question answering, semantic similarity, and text classification.

2. **Bag of Words - BoW:** Bag of Words (BoW) is a basic method for extracting meaningful features from text data. In this method, word frequencies in the text are counted and a word frequency function is constructed. Each word is considered as a feature, and the amount of these words in the text is the

representation of these features. BoW has the advantages of being simple and fast, but has the disadvantage that it does not take into account context and word order.

3. **TF-IDF:** TF-IDF is a method for extracting more comprehensive features from text data. In this method, the ratio of the frequency of a word in a document (term frequency) to its prevalence in the document (inverse document frequency) is calculated. In this way, uncommon but important terms are given more weight. Although TF-IDF offers the advantage of more information, it ignores context and word order.

4. **Word Embeddings:** Word Embeddings are dense, low-dimensional vector representations of words. Semantic links between words are captured by comparing these vectors. Examples of word embedding models are Word2Vec, GloVe and FastText. Although word embedding has the advantage of capturing semantic links, it does not fully reflect the context of the word in the sentence because it creates a fixed vector for each word.

5. **Other Transformer Based Models:** Models such as GPT, RoBERTa and T5 are examples of other transformer based models. These models also provide successful results in various NLP tasks; however, they may require high computational resources.

5.1.2.2 Audio Feature Extraction

Audio feature extraction is known as the process of taking raw audio signals and converting them into meaningful, processable features. These features can be used in tasks such as speech recognition, sentiment analysis and sound classification. Audio features represent the attributes of sound. Some audio feature extraction methods are as follows:

1. **OpenSMILE (Speech & Music Interpretation by Large-space Extraction):** OpenSMILE [15] is an open source tool widely used in audio feature extraction and analysis. It is frequently used in technical and research applications. OpenSMILE is a powerful software library that can extract various features from audio signals. Working principle of OpenSMILE:

   - **Feature Extraction:** OpenSMILE can extract multiple attributes to represent the characteristics and structure of audio signals. Examples of these attributes are energy, frequency spectrum, time and frequency attributes.

- **SES/LIW Segmentation:** Using OpenSMILE, audio signals can be separated into word and language segments. These segments can be used to study various properties of various audio signal components. For example, separating speech segments and extracting their features is used for speech recognition.
- **Audio Signal Normalization:** OpenSMILE normalises input audio signals to eliminate problems caused by audio signals from different sources being at different heights. The normalisation process is done to balance the overall dynamic range of the audio signal.
- **Preprocessing and Filtering:** OpenSMILE can take the audio signal through pre-processing stages such as removing the envelope, applying high-pass or low-pass filters, and applying noise reduction techniques.These pre-processing techniques help to reduce or eliminate unwanted elements from the signal.

2. **Wav2Vec:**Wav2Vec [16] is a model developed by Facebook that is mainly used in speech recognition tasks. Wav2Vec extracts learned feature representatives before directly converting sound waves into text labels based on unsupervised learning principles. These extracted features can then be used as input to traditional machine learning models.

3. **Librosa:** Librosa [17] is a Python-based library specifically used for music and sound analysis. It has capabilities to extract various features such as mel-frequency cepstrum coefficients (MFCCs), tempogram, chroma features. It can also perform audio data loading, sampling, and other pre-processing tasks.

4. **YAMNet (Yet Another MusicNet):**Developed by Google, YAMNet [18] is a pre-trained deep learning model. YAMNet is used to extract various features from audio data. By directly using these extracted features, tasks such as music classification can be performed.

5. **PyAudioAnalysis:** PyAudioAnalysis [19] is a Python library and is used for audio analysis, audio mining and other audio-based tasks. This library has the capabilities to extract many features such as mel-frequency cepstrum coefficients, energy, rhythm features.

5.1.2.3 Video Feature Extraction

Video feature extraction is the process of extracting both meaningful and processable features from cropped video segments. These features allow the capture of visual cues such as gestures and facial expressions in the video. There are various methods used for video feature extraction. Some of these methods are as follows:

1. **Convolutional Neural Networks (CNNs):** By dividing the video into consecutive frames, CNN examines each frame as a separate image to extract the video's features. The spatial feature vectors are inspected by a pre-trained CNN model (e.g. ResNet) and then extracted from each frame. By combining these feature vectors, the overall features of the video are represented. These feature vectors can be analyzed using models capable of sequential analysis when it is desired to extract temporal information from video. This structure efficiently extracts and analyzes both the temporal and spatial properties of video.

2. **3D Convolutional Neural Networks (3D CNNs):** 3D-CNN architecture, concurrently processes the temporal and spatial elements to extract information from videos. 3D convolution layers extract video features in three dimensions - width, height and time - by applying them to consecutive blocks of frames in the video. The interactions between movement and spatial patterns in video clips are captured in this manner. The resulting 3D feature maps produce robust representations of spatial and temporal information in the video data. This technique is known to be widely used in tasks such as event detection and motion recognition to extract comprehensive and in-depth features for video analysis.

   The layers of 3D-CNN are as follows:

   - **Convolutional Layer:** Applies a couple of filter kernels to the input. The input layer allows to reduce the complexity of the model by reducing the number of free parameters and weight sharing between pixels in the data stack. Filters act as feature detectors and thus identify specific features in the data. After shifting each convolution through a window of one kernel size, these filters provide a feature map. Pre-trained filters highlight certain aspects of the incoming data.
   - **Pooling Layer:** A summarization is performed on the subsamples of each feature map using techniques such as max pooling and average pooling. As a result, both the computational cost and the chance of overfitting are reduced.
   - **Normalization Layer:** Some normalization methods (e.g. Batch Normalization) normalize the output of layers and speed up the training process. Another advantage of this process is that it prevents overfitting.
   - **Activation Layer:** This layer increases the non-linearity of the output of the upper layers.

    - **Fully Connected Layer:** This layer flattens the features to produce the final output, which is then sent to the final output layers such as regression or classification.

3. **Optical Flow:** The optical flow extracts video features by calculating the motion of individual pixels in subsequent video frames. During this process, the pixel movements in each frame are represented as a vector. The resulting optical flow vectors can then be used as features to determine the direction and amount of motion. The motion of objects in the video is captured and tracked using optical flow vectors. This technique works well in problems where dynamic changes and movements in videos need to be captured.

4. **Pretrained Models and Transfer Learning:** This technique extracts video features by utilizing models that have been previously trained using existing large datasets. A pre-trained model (for example, a ResNet trained on CIFAR or a CNN trained on ImageNet) extracts features from each frame. Transfer learning is used to adjust the weights of the pre-trained model so that the model performs effectively on a new dataset. This approach ensures good performance even on smaller datasets and saves time when analyzing videos.

5.1.3 Dimensionality Reduction

The process of reducing the size of high-dimensional datasets to make them easier to analyze is known as dimensionality reduction.

5.1.3.1 Principal Componenyt Analysis (PCA)

6373 different features are obtained as a result of feature extraction with OpenSMILE on audios. The 6373 different features obtained with OpenSMILE provide a complex representation of audio data. This number of features is often high and this large data set may contain a large number of redundant or correlated features beyond a few important nuances. This can cause a number of problems and necessitates the use of dimensionality reduction techniques such as Principal Component Analysis (PCA).

It is known that high-dimensional datasets often complicate the learning process of the model. They can also increase the risk of overfitting. In such cases, reducing the number of features can help the model to better generalize the data and produce more robust results. Dimensionality reduction techniques, such as PCA, can address this problem by transforming the representation of the data into a lower dimensional subspace.

Within the scope of the project, 6373 features were reduced to 100 features by applying PCA.

5.1.3.2 Max Pooling

Another method used to reduce dimensionality in deep learning models is max pooling. Below is a description of its working mechanism:

- Maximum pooling applies a pooling window (filter), usually square shaped, on the input.
- The window is scrolled across the image with a specific stride, which defines the stride size by which the window is scrolled.
- The highest value in the window is selected at each position and stored as pooled output.
- The outcome is a reduced-size map that keeps the most important features after this method is applied to the complete feature map.

5.1.4 Modelling

Model architectures used in this study are explained in this section.

5.1.4.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is Google's neural network-based pre-training method for natural language processing. Google uses this method to provide users with more accurate results by better analyzing their search queries.

BERT evaluates the sentence both left-to-right and right-to-left, in contrast to other models. The basis for this model is transformer architecture. By examining the words that come before and after a word, text can determine its context. The relationships between words and their meanings are better determined in this way.

Unlike other models, BERT evaluates a sentence from both right-to-left and left-to-right. This model comes from transformer architecture. The context of a text can be understood by examining the words that come before and after a word. In this way, the relationships between the meanings of words can be better understood.

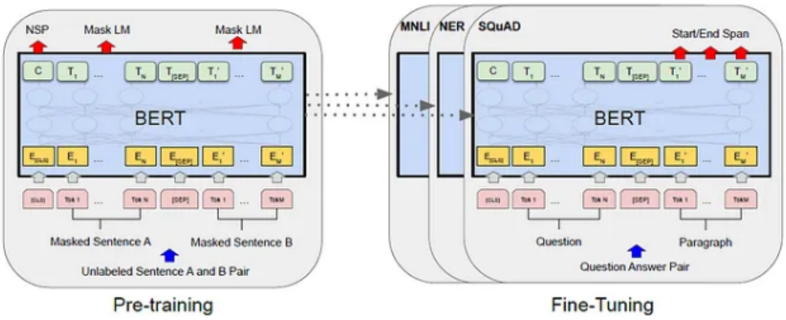


Figure 5.2 BERT Architecture [13]

5.1.4.2 XGBoost (Extreme Gradient Boosting):

XGBoost is a gradient boosting library designed to give machine learning algorithms better accuracy and efficiency.

As shown in Figure 5.3, by combining a large number of simple models or 'weak predictors' such as decision trees, the Gradient Boosting framework produces a strong predictor. Each new model is trained to reduce the errors of the previous weak predictors. Furthermore, in order to prevent overfitting, it incorporates a variety of regularization approaches, including L1 and L2 Regularization. This keeps the model from overfitting and improves its generalizability.

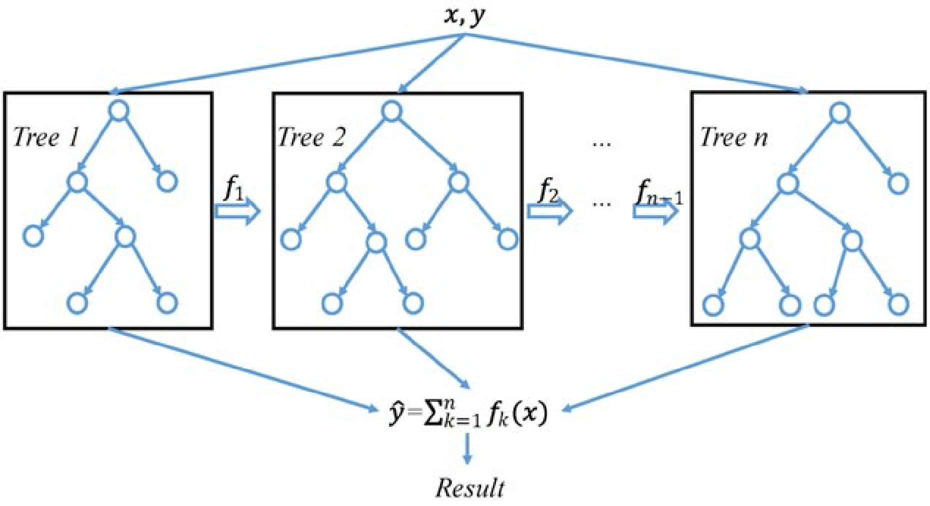


Figure 5.3 A General Architecture of XGBoost [20]

5.1.4.3 Bidirectional Long Short-Term Memory (BiLSTM):

The BiLSTM architecture consists of a bidirectional LSTM that processes the sequence both forward and backward directions. This architecture can be thought of as two independent LSTM networks, one receiving the token sequence in its original order and the other in reverse order. The combined probability obtained from the outputs of these two LSTM networks is the final output. Each of these networks produces a probability vector as output [21]. It can be shown as in Figure 5.4.

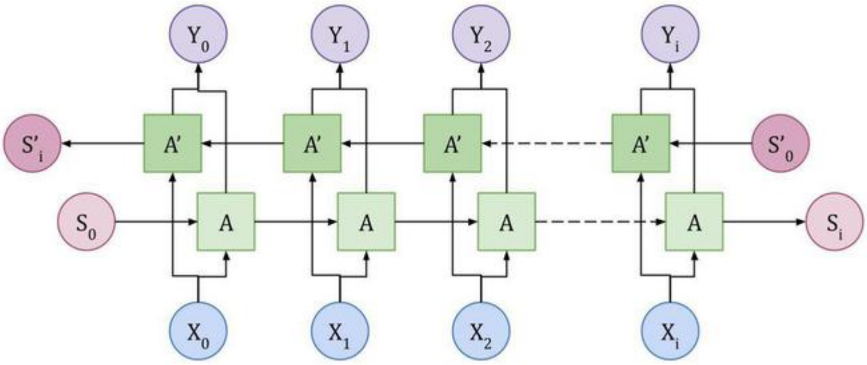


Figure 5.4 A General Architecture of BiLSTM [21]

Table 5.1 visualizes which of the models described above are used as classifiers with which selected features.

	BERT	BiLSTM	XGBoost
Text Model	X	X	-
Audio Model	-	X	X
Video Model	-	X	X
Combined Models	-	-	X

Table 5.1 Models Used in Response to Selected Features

6
Application

The characteristics and training of the models will be covered in this chapter.

6.1 Extracting Features

3 different methods are used to extract different type of data.

- **Text Feature Extraction** BERT (Bidirectional Encoder Representations from Transformers) is used to extract text features. Modern transformer-based models such as BERT take into account the words that come before and after a word to understand its context within a sentence. BERT is capable of creating highly contextualized word embeddings that can accurately represent the subtleties and complexity of natural language. These embeddings serve as rich feature representations for tasks such as entity recognition, sentiment analysis, and text classification.

- **Audio Feature Extraction** OpenSMILE (Open Source Media Interpretation with Large Feature Space Extraction) is a flexible toolset for extracting features from audio signals. It can process a wide range of acoustic features, including low-level descriptors such as pitch and energy, as well as high-level features associated with speech and emotion. OpenSMILE is particularly useful for tasks such as speaker identification, emotion detection, and speech recognition.

- **Video Feature Extraction** In this study, 3D-CNN is preferred for video feature extraction task. 3D-CNN is a good candidate for feature extraction from videos and data containing temporal information (because it examines the data in three dimensions). They have the capacity to extract more comprehensive features from videos, taking into account both temporal information and two-dimensional visual information. So, 3D-CNN design is ideal for examining how items move, behave, and interact with one another.

6.2 Splitting the Dataset into Train, Test, Validation

SEMEVAL2024 Task 3: There are a total of 13587 dialogues in the dataset. The data set is divided into 70% (9782) for train, 20% (2446) for testing, and 10% (1359) for validation.

IEMOCAP: There are a total of 7433 dialogues in the dataset. The data set is divided into 70% (5202) for train, 20% (1487) for testing, and 10% (744) for validation.

6.3 Combining Features

By using the concatenate method of the numpy library all three multimodal features, the final representation of an utterance u is created.

$$\mathbf{f} = \mathbf{t} \oplus \mathbf{a} \oplus \mathbf{v}$$

The model will be fed with these successively added data.

6.4 Training the Models

This section describes the parameters and values used by the models.

6.4.1 BiLSTM

- **embed_len:** The length of each word's vector representation. This parameter is given to the first layer, the Embedding layer
- **units=80:** Number of hidden units of each LSTM cell. This parameter is given to LSTM layers.
- **return_sequences=True:** Utilized at every time step in the sequence to extract characteristics. Learning complicated time dependencies is aided by the consideration of all time steps. This parameter is given to LSTM layers.
- **GlobalMaxPool1D():** By taking the highest value of the time steps, this layer generates an output with a fixed size. This aids in the model's selection of the most notable features. This layer is added after the BiLSTM layers.
- **Dropout(rate=0.5):** At every training stage, 50% of the neurons are randomly deactivated. Dropout inhibits overfitting and improves the model's capacity for generalization.

- **activation="relu":** In neural networks, activation functions are used to calculate the output of each neuron and add non-linearity to the model. ReLU activation function, returns the result exactly as it is if the input value is positive; if it is negative, it sets the output to zero.

- **activation="softmax":** Softmax activation function, transforms a set of values into a probability distribution. Every value has a result that falls between 0 and 1.

6.4.2 BERT

- **Optimizer:** The model's parameters are optimized using an AdamW optimizer. Weight decay is the method of weight adjustment used by AdamW, a derivative of the Adam optimizer. The configuration of the optimizer involves configuring the epsilon value to 1e-8 and the learning rate to 2e-5.

- **Learning Rate Scheduler:** A learning rate scheduler is developed that, depending on the optimizer, increases the learning rate over a warm-up period of time and then decays at a linear pace. In the process of training a model, this is used to maximize learning rate.

- **Number of Epochs:** Training and validation cycles are initiated for 20 epochs.

- **Batch Size:** By utilizing every training batch during the training cycle, it enables the model to operate in training mode. Next, it passes forward, computes the loss, computes the backward gradients, resets the previously computed gradients for each batch, and updates the parameters. The batch size is 32.

- **Early Stopping:** If the validation loss begins to rise, an early stopping mechanism is used to halt training. This indicates that overfitting has started and that the model's performance is no longer increasing.

6.4.3 XGBoost

- **objective='multi:softmax':** This defines the objective function to be applied to the multi-class classification problem of XGBoost. "Multi" denotes an objective for multi-class classification that makes use of the softmax function.

- **num_class:** This indicates how many classes the model is probably going to predict.

- **max_depth:** This indicates each decision tree's maximum depth. Greater values could result in more intricate models, but they could also raise the possibility of overfitting. It was found to be 5.

- **learning_rate:** This hyperparameter regulates the model's learning rate at each learning step. While a low learning rate might make training more consistent and dependable, it might also make learning more slowly. It was found to be 0.1.

- **n_estimators:** This indicates how many trees are to be used. XGBoost combines several weak estimators (usually decision trees) to produce a strong estimator. The number of these poor estimators to be used is indicated by *n_estimators*. Although a higher value may raise the possibility of overfitting, it can also result in a more complex model. It was found to be 100.

7
Experimental Results

The project's experimental outcomes are discussed in this chapter.

7.1 Model Results

In single models, BERT & Keras was used as text feature extractor and BERT for BERT features, BiLSTM for Keras features as a classifier for text classification; openSMILE was used as audio feature extractor and XGBoost & BiLSTM as classifier for audio classification; 3D-CNN was used as feature extractor and XGBoost & BiLSTM as classifier for video classification.

The fact that BiLSTM is a more traditional method resulted in its inability to produce high scores against the complex and noisy data used in this study. Furthermore, imbalance in the dataset is also a challenge for the BiLSTM architecture as it performs poorly in classes with a small number of instances. Table 7.1 shows the success of BiLSTM versus the other models used on IEMOCAP and SemEval datasets.

Another point is that BERT has achieved better performance than Keras as a text feature extractor.

		SemEval	**IEMOCAP**	**SemEval**	**IEMOCAP**
	Model	**Accuracy**		**Macro Avg. F-1**	
Text	BiLSTM	0.26	0.46	0.43	0.39
	BERT	0.64	0.59	0.47	0.57
Audio	BiLSTM	0.37	0.40	0.11	0.36
	XGBoost	0.41	0.38	0.12	0.35
Video	BiLSTM	0.44	024	0.09	0.13
	XGBoost	0.43	0.27	0.10	0.26

Table 7.1 Model Comparison

For this reason, the outputs given for single models in the following sections are scores from the BERT classifier and feature extractor for text and XGBoost for audio and video.

In combined models, previously extracted features were combined before being fed into the model and XGBoost was used as the classifier. Model outputs were obtained in this way.

When comparing model outputs, mostly the F-1 macro averages are considered rather than accuracy and weighted average.

7.1.1 Single Models

The features were given to the models individually before being combined and the outputs were obtained. The purpose of this is to create a baseline to see how the combination affects the emotion detection.

Table 7.2 shows the scores for SemEval and IEMOCAP datasets. As mentioned earlier, SemEval is a highly imbalanced dataset, so it was expected that IEMOCAP dataset would produce higher scores.

	Dataset	**Accuracy**	**Macro Avg. F-1**	**Weighted Avg. F-1**
Text	SemEval	0.64	0.47	0.63
	IEMOCAP	0.59	0.58	0.59
Audio	SemEval	0.41	0.12	0.29
	IEMOCAP	0.38	0.35	0.37
Video	SemEval	0.43	0.09	0.27
	IEMOCAP	0.27	0.26	0.27

Table 7.2 Single Model Outputs

7.1.2 Combined Models

Features are added one after the other using the concatenation method when combining features. Outputs of the combined models is shown in Table 7.3. As with the single models, the score dominance of the IEMOCAP dataset over SemEval is clearly seen here.

Looking at the combined model outputs, it is seen that the model with text, audio and video features has the highest score. Possible reasons for this will be discussed in the Performance Analysis Section (8).

Combined Features	Dataset	Accuracy	Macro Avg. F-1	Weighted Avg. F-1
Text + Audio	SemEval	0.49	0.20	0.40
	IEMOCAP	0.47	0.44	0.46
Text + Video	SemEval	0.48	0.18	0.38
	IEMOCAP	0.48	0.46	0.47
Audio + Video	SemEval	0.44	0.09	0.27
	IEMOCAP	0.54	0.51	0.53
Text+Audio+Video	SemEval	0.48	0.18	0.37
	IEMOCAP	0.55	0.53	0.54

Table 7.3 Combined Model Outputs

The reasons why IEMOCAP data set produces better results than SemEval data set are the following:

- As seen in Figure 1.2 SemEval data set contains a high level of imbalance. There are a total of 13587 data, of which 5929 are in a single class, neutral. That means neutral class covers 43.53% of total data. On account of that reason classification models fed with SemEval data are so prone to predict neutral class and this decreases the accuracy significantly.
- In the IEMOCAP video data, there is one person with clearly visible faces and hands and a clean background. In SemEval, on the other hand, there are many irrelevant objects in the background. Also, the faces of the speakers are not always very clear.
- Since the SemEval dataset is based on a sitcom, emotions are less reflected in the tone of voice due to sarcastic scenes, whereas in IEMOCAP emotions are reflected in the tone of voice.

8
Performance Analysis

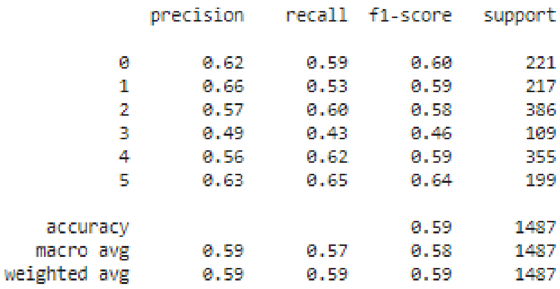
This chapter analyzes the performance of the highest scoring models of the outputs given in Table 7.2 and Table 7.3.

8.1 Single Models

Looking at Table 7.2, it can be seen that higher success was achieved in all of the text classification, audio classification and video classification models in the IEMOCAP data set. In this title, the outputs of these 3 models will be analyzed in detail.

8.1.1 Text Classification Model

Figure 8.1 shows the scores obtained on a class basis for BERT text classification model trained on IEMOCAP dataset. Numeric values of the labels: ['ang': 0, 'exc': 1, 'fru': 2, 'hap': 3, 'neu': 4, 'sad': 5].



```
              precision    recall  f1-score   support

           0       0.62      0.59      0.60       221
           1       0.66      0.53      0.59       217
           2       0.57      0.60      0.58       386
           3       0.49      0.43      0.46       109
           4       0.56      0.62      0.59       355
           5       0.63      0.65      0.64       199

    accuracy                           0.59      1487
   macro avg       0.59      0.57      0.58      1487
weighted avg       0.59      0.59      0.59      1487
```

Figure 8.1 BERT Model's Classification Report for IEMOCAP Text Data

It can be inferred from Figure 8.1 that although there is an imbalance between classes and also, based on the support metric on the classification report and Figure 1.3 it can be seen that the majority classes in the dataset are 'neu' and 'fru', the output shows that the majority class does not dominate the minority class and the model provides a robust result as the model does not overfit.

As the confusion matrix of the BERT text model in Figure 8.2 demonstrates, the model most frequently predicts the labels "fru" and "ang" interchangeably which corresponding to the emotions frustrated and angry in real life. The reason for this inaccuracy in the predictions can be interpreted as the fact that these two emotions have common points with each other and may contain similar verbal expressions.

For example, this sentence: "All right! But don't think like that, because I mean, what the hell did we do all this for, Chris? The whole business it's all for you! The whole shooting match is for you!" is labeled with "ang" and this sentence: "What have I got to hide? What the hell is the matter with you, Kate?" is labeled with "fru". As just mentioned above, the two sentences have the same expression which is "what the hell". This may have caused the model to mispredict.

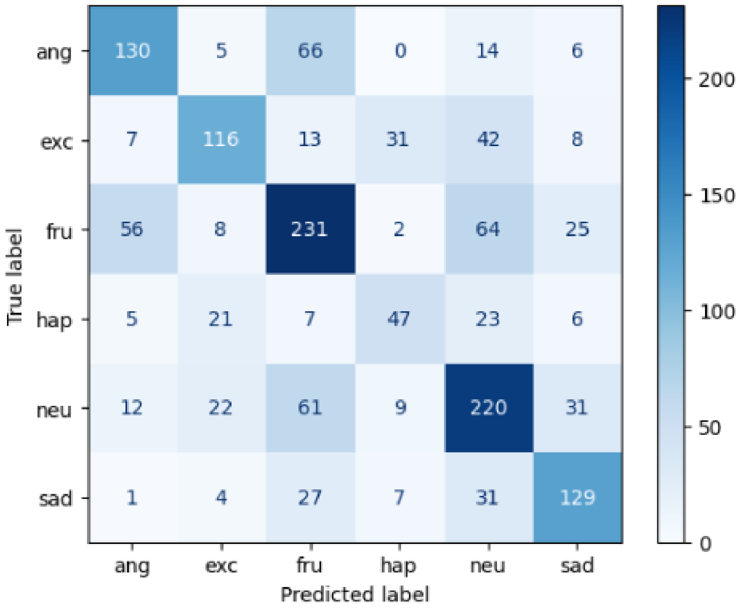
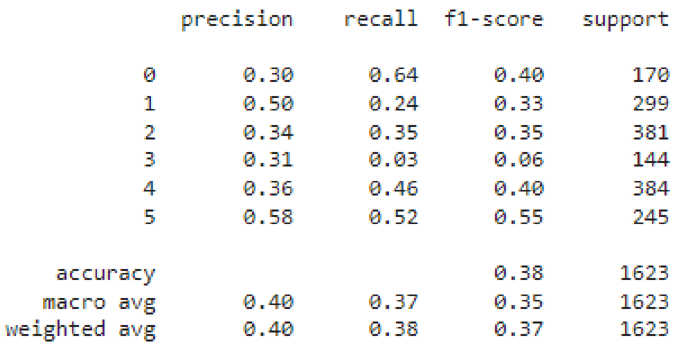


Figure 8.2 BERT Model's Confusion Matrix for IEMOCAP Text Data

8.1.2 Audio Classification Model

Figure 8.3 shows the scores obtained on a class basis for XGB audio classification model trained on IEMOCAP dataset. Numeric values of the labels: ['ang': 0, 'exc': 1, 'fru': 2, 'hap': 3, 'neu': 4, 'sad': 5].

Looking at Figure 8.3, it can be seen that the f1-score values of the 0th (anger) and 5th (sad) classes are high and the 3rd (happy) class is low. This is because it is easier to reflect the emotions of anger and sadness in the voice than to reflect happiness. Voice tone is more likely to be distinctive in speech that reflects angry or sad emotions.



```
              precision    recall  f1-score   support

           0       0.30      0.64      0.40       170
           1       0.50      0.24      0.33       299
           2       0.34      0.35      0.35       381
           3       0.31      0.03      0.06       144
           4       0.36      0.46      0.40       384
           5       0.58      0.52      0.55       245

    accuracy                           0.38      1623
   macro avg       0.40      0.37      0.35      1623
weighted avg       0.40      0.38      0.37      1623
```

Figure 8.3 XGB Model's Classification Report for IEMOCAP Audio Data

When we look at the confusion matrix of the XGB audio model in Figure 8.4, it is seen that the model generally tends to produce the 4th (neutral) class as a prediction. The model may have perceived these speeches as neutral because the speakers did not always fully reflect their emotions in their voices and the speeches sometimes lacked emphasis.

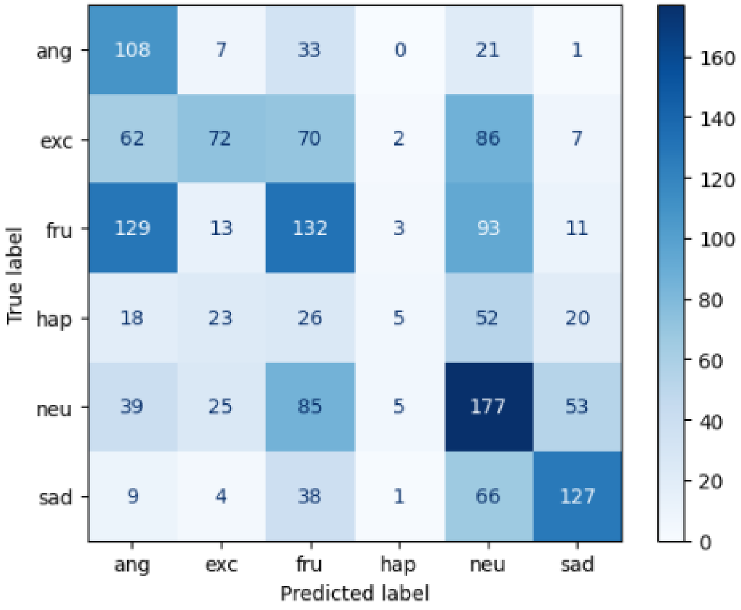
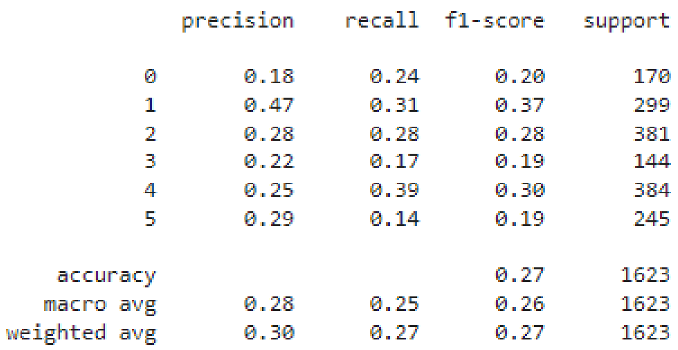


Figure 8.4 XGB Model's Confusion Matrix for IEMOCAP Audio Data

8.1.3 Video Classification Model

Figure 8.5 shows the scores obtained on a class basis for XGB video classification model trained on IEMOCAP dataset. Numeric values of the labels: ['ang': 0, 'exc': 1, 'fru': 2, 'hap': 3, 'neu': 4, 'sad': 5].



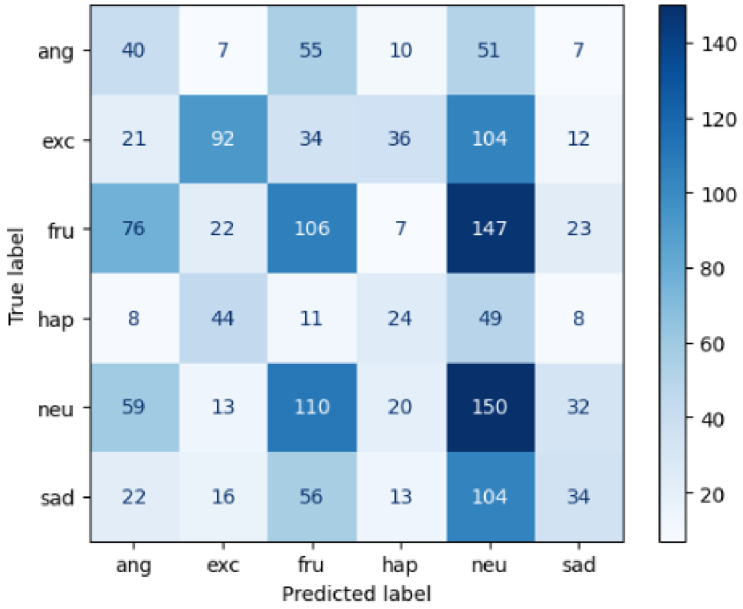
```
              precision    recall  f1-score   support

           0       0.18      0.24      0.20       170
           1       0.47      0.31      0.37       299
           2       0.28      0.28      0.28       381
           3       0.22      0.17      0.19       144
           4       0.25      0.39      0.30       384
           5       0.29      0.14      0.19       245

    accuracy                           0.27      1623
   macro avg       0.28      0.25      0.26      1623
weighted avg       0.30      0.27      0.27      1623
```

Figure 8.5 XGB Model's Classification Report for IEMOCAP Video Data

Looking at Figure 8.5, it is seen that the 1st (excited) class has the highest f1-score value. The reason for this is that in speeches that reflect excited emotions, speakers emphasize the emotion by making more gestures and facial expressions.

The reason why the 4th (neutral) class has a high f1-score is that the model tends to predict the 4th class, as seen in the confisuon matrix in Figure 8.6. As mentioned earlier, when speakers do not support their speech with distinctive gestures, facial expressions, or tone of voice, the model tends to be directly labeled as neutral.

Contrary to the relatively high f1-score of the 4th grade, the low precision value shows that most of the samples classified as positive by the model are actually false positives.



	ang	exc	fru	hap	neu	sad
ang	40	7	55	10	51	7
exc	21	92	34	36	104	12
fru	76	22	106	7	147	23
hap	8	44	11	24	49	8
neu	59	13	110	20	150	32
sad	22	16	56	13	104	34

Figure 8.6 XGB Model's Confusion Matrix for IEMOCAP Video Data

The reasons why text data produces better results than audio and video may be the following:

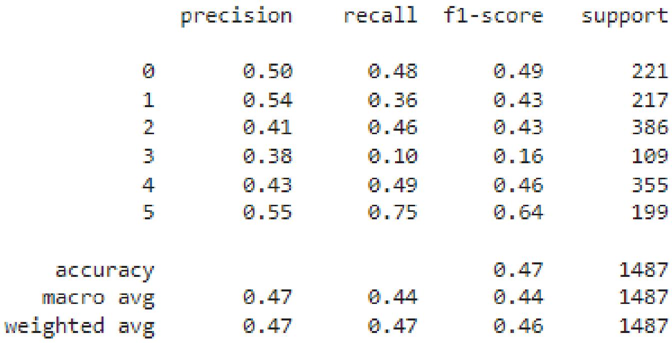
- Textual data typically has a direct and straightforward meaning. Whereas audio and video data types are more prone to noise. When it comes to audio, environmental elements like background noise and microphone quality might degrade the quality of the recorded data. For video, the same holds true. There are elements like illumination, resolution, etc.
- It is more difficult to extract high quality features from noisy data, and the lower the feature quality, the lower the model performance.
- Video data may contain non-emotional elements, e.g. a static object in the scene, etc.

8.2 Combined Models

Looking at Table 7.3, it can be seen that higher success was achieved in all of the combined models in the IEMOCAP data set. In this title, the outputs of these 4 models will be analyzed in detail.

8.2.1 Text - Audio Classification Model

The classification report of the text-audio combined model is seen in Figure 8.7. When the F-1 scores are analyzed in the classification report, the poor score in class 3, which is happy label, draws attention. One possible reason for this is that the text and audio features may have been inadequate to express the emotion of happiness because it may not be as easy to extract from the audio as emotions such as anger. A person's happiness can largely be understood from their facial expressions and gestures. Therefore, in this combination, class 3 is less predicted by the model than the others.



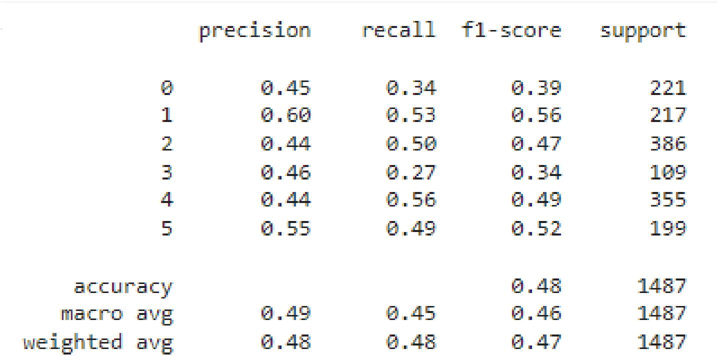
```
              precision    recall  f1-score   support

           0       0.50      0.48      0.49       221
           1       0.54      0.36      0.43       217
           2       0.41      0.46      0.43       386
           3       0.38      0.10      0.16       109
           4       0.43      0.49      0.46       355
           5       0.55      0.75      0.64       199

    accuracy                           0.47      1487
   macro avg       0.47      0.44      0.44      1487
weighted avg       0.47      0.47      0.46      1487
```

Figure 8.7 Classification Report of Text-Audio Combined Model

8.2.2 Text - Video Classification Model

Figure 8.5 shows the scores obtained on a class basis for XGB text-video classification model trained on IEMOCAP dataset. Numeric values of the labels: ['ang': 0, 'exc': 1, 'fru': 2, 'hap': 3, 'neu': 4, 'sad': 5].

When the Figure 8.8 is examined it is seen that class 0 , which is anger, has the lowest F-1 score when it is compared with the other combined models on the basis of class 0. The inference that can be drawn from this is that audio features are of great importance in the prediction of anger emotion.



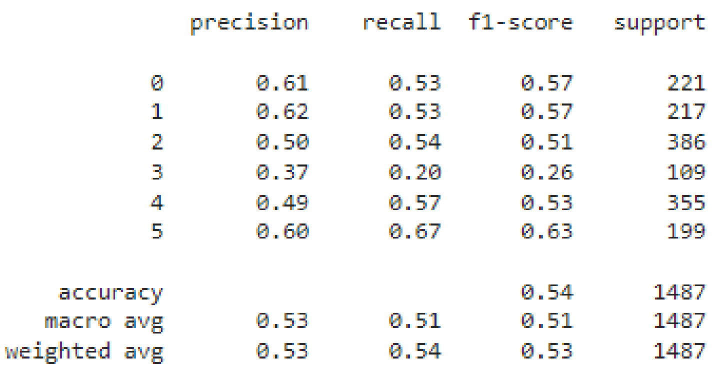
```
              precision    recall  f1-score   support

           0       0.45      0.34      0.39       221
           1       0.60      0.53      0.56       217
           2       0.44      0.50      0.47       386
           3       0.46      0.27      0.34       109
           4       0.44      0.56      0.49       355
           5       0.55      0.49      0.52       199

    accuracy                           0.48      1487
   macro avg       0.49      0.45      0.46      1487
weighted avg       0.48      0.48      0.47      1487
```

Figure 8.8 Classification Report of Text-Video Combined Model

8.2.3 Audio - Video Classification Model

Figure 8.5 shows the scores obtained on a class basis for XGB audio-video classification model trained on IEMOCAP dataset. Numeric values of the labels: ['ang': 0, 'exc': 1, 'fru': 2, 'hap': 3, 'neu': 4, 'sad': 5].

When the classification report in Figure 8.9 is analyzed, it can be said that the audio and video features are complementary to each other because it produced a very close result with the text-audio-video combined model which gave the highest success. Also, unlike the models using text-audio and text-video features, there is no significant F-1 score reduction in class basis.

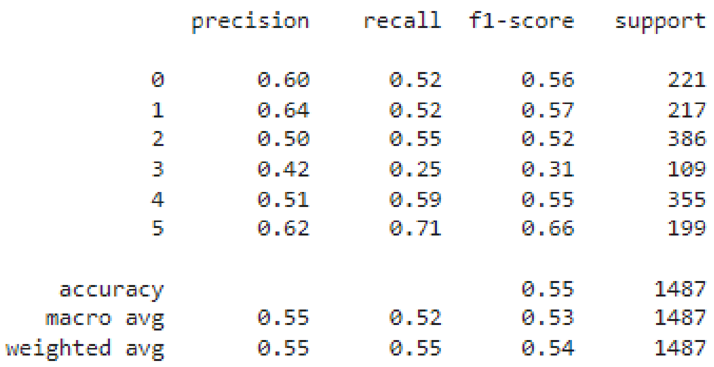


```
              precision    recall  f1-score   support

           0       0.61      0.53      0.57       221
           1       0.62      0.53      0.57       217
           2       0.50      0.54      0.51       386
           3       0.37      0.20      0.26       109
           4       0.49      0.57      0.53       355
           5       0.60      0.67      0.63       199

    accuracy                           0.54      1487
   macro avg       0.53      0.51      0.51      1487
weighted avg       0.53      0.54      0.53      1487
```

Figure 8.9 Classification Report of Audio-Video Combined Model

8.2.4 Text - Audio - Video Classification Model

It is seen that the text-audio-video combined model gives the best result among the other combined models. Figure 8.10 shows the classification report of the model. It can be inferred that model does not lead predictions to focus on one class or a group of majority classes.



```
              precision    recall  f1-score   support

           0       0.60      0.52      0.56       221
           1       0.64      0.52      0.57       217
           2       0.50      0.55      0.52       386
           3       0.42      0.25      0.31       109
           4       0.51      0.59      0.55       355
           5       0.62      0.71      0.66       199

    accuracy                           0.55      1487
   macro avg       0.55      0.52      0.53      1487
weighted avg       0.55      0.55      0.54      1487
```

Figure 8.10 Classification Report of Text-Audio-Video Combined Model

When the confusion matrix in the Figure 9.2 is examined, it is seen that the same situation is faced. The model most frequently predicts the labels 'fru' and 'ang' interchangeably. 'neu' and 'fru' labels are also predicted interchangeably. However, there is a very low level of interchangeability between other labels.

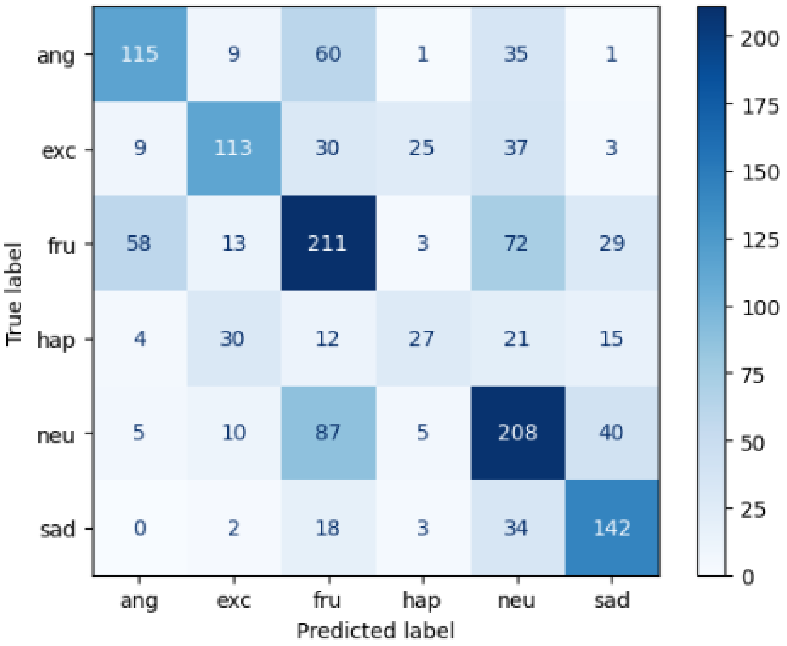


Figure 8.11 Confusion Matrix of Text-Audio-Video Combined Model

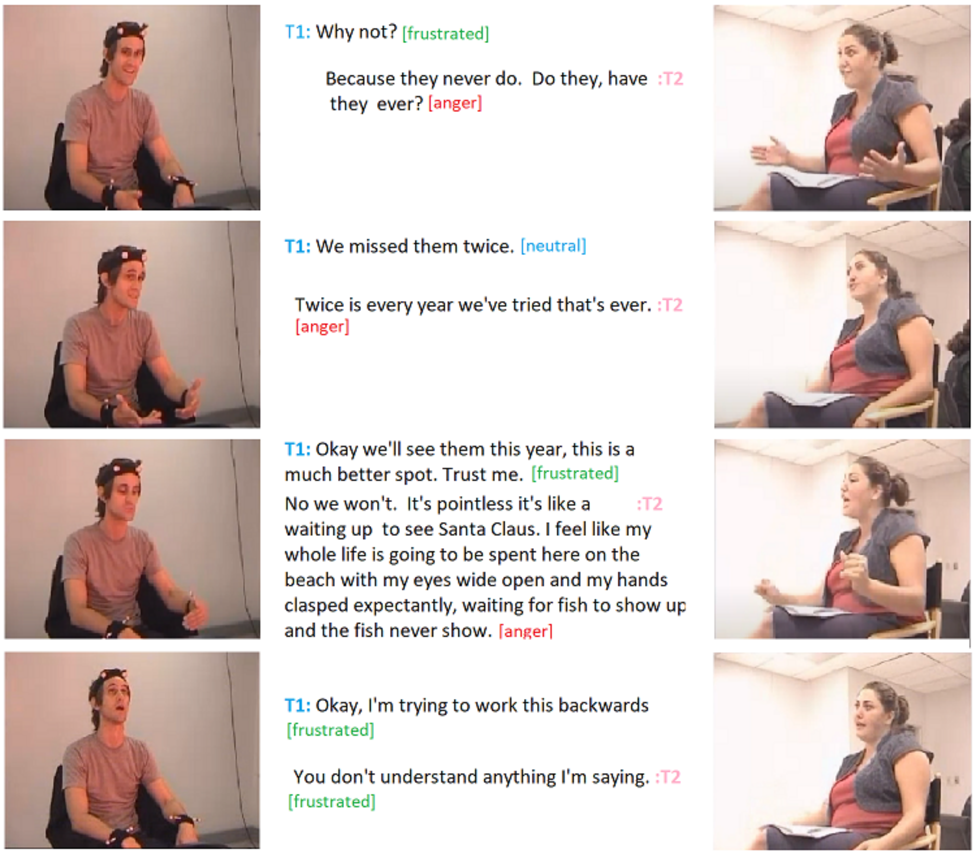
Looking at the combined models, combining text embeddings with audio or video features positively affects success. This is a logical conclusion, given that the single text model has the highest success. The final result of this study is that the combination of text, audio and video has the highest score among the combined models. Instead of using single models for audio and video for emotion detection, combining all three types of data produces better results for this study. It can be stated that this result is robust because the study was tested on two distinct datasets, and the same circumstances applied to both.

9
Result

This chapter shows a sample dialogue with its predicted emotions and provides a general evaluation.

9.1 Sample Dialogue

In this section, the emotion predictions made by the text classification model, which has achieved the highest success in single models, and the text-audio-video classification model, which has achieved the highest success in combined models, for scenes in a dialogue are given.



T1: Why not? [frustrated]

Because they never do. Do they, have :T2
they ever? [anger]

T1: We missed them twice. [neutral]

Twice is every year we've tried that's ever. :T2
[anger]

T1: Okay we'll see them this year, this is a much better spot. Trust me. [frustrated]

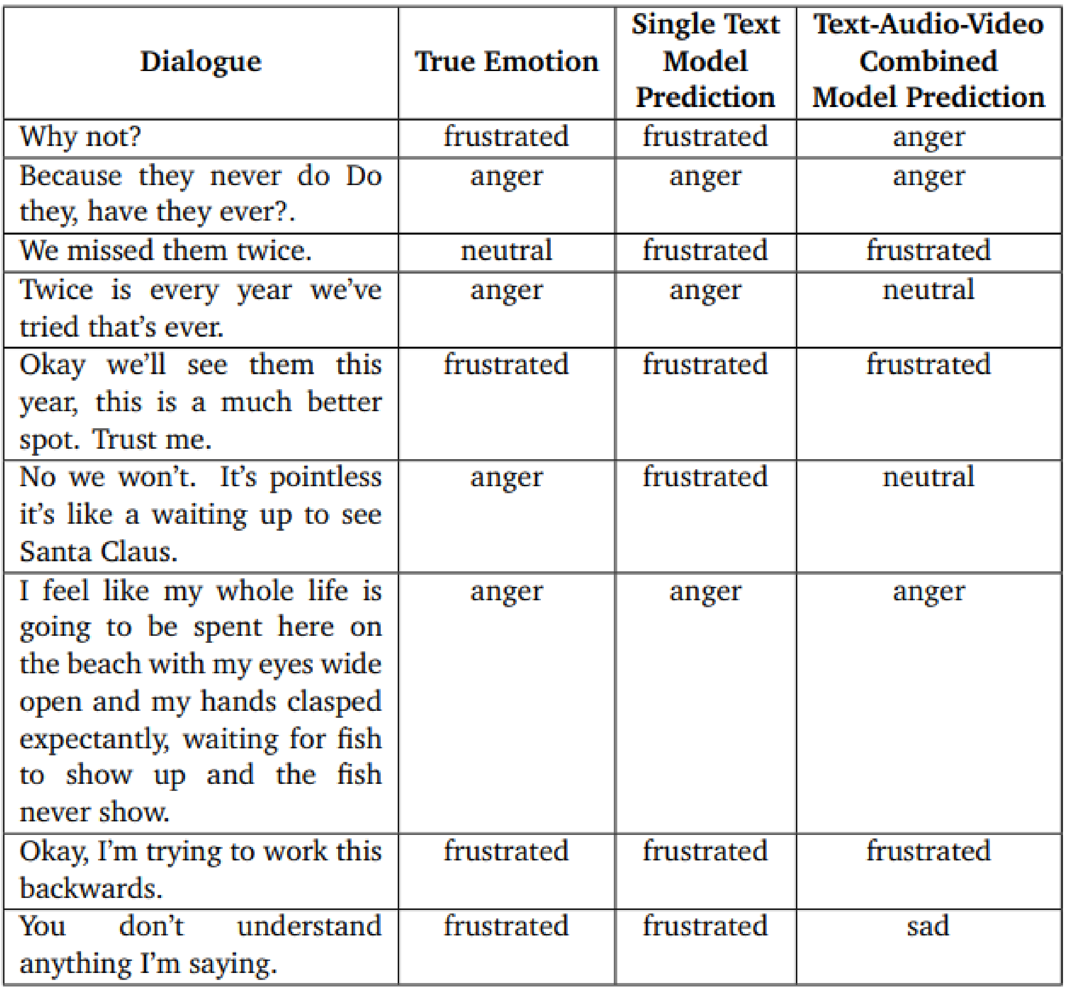
No we won't. It's pointless it's like a :T2
waiting up to see Santa Claus. I feel like my whole life is going to be spent here on the beach with my eyes wide open and my hands clasped expectantly, waiting for fish to show up and the fish never show. [anger]

T1: Okay, I'm trying to work this backwards [frustrated]

You don't understand anything I'm saying. :T2
[frustrated]

Figure 9.1 Dialogue Sample

For the dialogue example given in Figure 9.1, predictions were obtained from the text model, which is the most successful of the single models, and the text-audio-video model, which is the most successful of the combined models.



Dialogue	True Emotion	Single Text Model Prediction	Text-Audio-Video Combined Model Prediction
Why not?	frustrated	frustrated	anger
Because they never do Do they, have they ever?.	anger	anger	anger
We missed them twice.	neutral	frustrated	frustrated
Twice is every year we've tried that's ever.	anger	anger	neutral
Okay we'll see them this year, this is a much better spot. Trust me.	frustrated	frustrated	frustrated
No we won't. It's pointless it's like a waiting up to see Santa Claus.	anger	frustrated	neutral
I feel like my whole life is going to be spent here on the beach with my eyes wide open and my hands clasped expectantly, waiting for fish to show up and the fish never show.	anger	anger	anger
Okay, I'm trying to work this backwards.	frustrated	frustrated	frustrated
You don't understand anything I'm saying.	frustrated	frustrated	sad

Figure 9.2 Sample Dialogue Emotion Predictions

It can be said that the text model makes slightly better emotion prediction than the text-audio-video combined model.

9.2 Conclusion

In this study, emotion detection from text, audio and video features is discussed. Within the scope of the study, text, audio and video features were combined in different ways and their effect on emotion detection was observed. Training and performance evaluation were conducted using two separate data sets: SemEval and IEMOCAP. As a result, it was possible to compare the data sets in terms of content.

As the first step of the study, the features of the text, audio and video data in the data sets used were extracted. Feature extraction was performed with BERT for text, OpenSMILE for audio, and 3D-CNN for video, and the work continued with these features.

To classify the extracted features, results were obtained using BERT model in text and XGB model in audio and video. According to the results shown by single models, using text for emotion detection was more successful than using audio and video.

Models were retrained by combining the extracted features in different ways. According to the results shown by combined models, combining audio and video features with text features has a positive effect on success. In combined models, the most successful results were obtained in the text-audio-video combined model.

Although the text-audio-video combined model is the most successful model among the combined models, it could not surpass the success of the text model. The reason for this may be that the gestures, facial expressions and tones of voice in some videos and audios contain misleading information. This misleading information may have negatively affected the model's prediction.

The following domains benefit from this study's outputs.

- **Security and Law:** By analyzing videos recorded from security cameras or forensic interrogations, emotion detection is crucial because it can help predict threats and crimes.
- **Workplace and Human Resources:** With emotion detection, motivation and satisfaction of employees in a workplace can be monitored and inclusive analysis can be made.
- **Customer Service and Call Centers:** By better understanding their customers' emotional states, agents can provide a more efficient service. For example, an angry customer who is dissatisfied with the service they have received can be dealt with more carefully and waiting times can be reduced.
- **Marketing and Advertising:** If it is known which ads evoke more positive emotional reactions from viewers, more accurate strategies can be developed in future advertising campaigns, taking into account positive or negative reactions.
- **Entertainment and Gaming:** The entertainment industry can use emotion detection to understand the impact that movies or video games have on audiences. This can lead to more satisfying content.

Combining models may be more appropriate in real-time applications, as using text, audio, or video alone might not yield successful results due to noisy and corrupted data.

References

[1] A. E. ÖZMUTLU, “Doğal dil işleme,” *Bilgisayar Bilimlerinde Teorik Ve Uygulamalı Araştırmalar*, vol. 129, 2021.

[2] (), [Online]. Available: https://kili-technology.com/data-labeling/nlp/mastering-entity-recognition-and-text-classification-for-machine-learning-engineers (visited on 11/30/2023).

[3] (), [Online]. Available: https://www.datarobot.com/blog/text-processing-what-why-and-how/ (visited on 11/29/2023).

[4] (), [Online]. Available: https://www.simplilearn.com/image-processing-article#what_is_image_processing (visited on 11/29/2023).

[5] (), [Online]. Available: https://www.v7labs.com/blog/image-processing-guide (visited on 11/29/2023).

[6] (), [Online]. Available: https://www.linkedin.com/pulse/introduction-image-processing-g%C3%B6r%C3%BCnt%C3%BC-i%C5%9Flemeye-giri%C5%9F-udentify/?originalSubdomain=tr (visited on 11/29/2023).

[7] (), [Online]. Available: https://dac.digital/audio-processing/ (visited on 11/29/2023).

[8] J. C. John R. Absher, *Neuroimaging Personality, Social Cognition, and Character*. 2016.

[9] (), [Online]. Available: https://semeval.github.io/SemEval2024/tasks.html (visited on 06/02/2024).

[10] C. Busso *et al.*, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[11] (), [Online]. Available: https://drive.google.com/drive/folders/1TIRBiL8z4ZnoxtuKM8pnjtm2BxB5mS4Y (visited on 11/29/2023).

[12] (), [Online]. Available: https://drive.google.com/file/d/1zWCN2oMdibFkOkgwMG2m02uZmSmynw8c/view (visited on 06/03/2024).

[13] (), [Online]. Available: https://medium.com/@ktoprakucar/bert-modeli-ile-t%C3%BCrk%C3%A7e-metinlerde-s%C4%B1n%C4%B1fland%C4%B1rma-yapmak-260f15a65611 (visited on 12/02/2023).

[14] (), [Online]. Available: https://ayselaydin.medium.com/1-text-preprocessing-techniques-for-nlp-37544483c007 (visited on 12/02/2023).

[15] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[16] (), [Online]. Available: https://huggingface.co/docs/transformers/model_doc/wav2vec2 (visited on 12/02/2023).

[17] (), [Online]. Available: https://librosa.org/doc/main/index.html (visited on 12/02/2023).

[18] (), [Online]. Available: https://www.kaggle.com/models/google/yamnet/frameworks/tensorFlow2/variations/yamnet/versions/1?tfhub-redirect=true (visited on 12/02/2023).

[19] (), [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis (visited on 12/02/2023).

[20] (), [Online]. Available: https://www.researchgate.net/publication/335483097_A_hybrid_ensemble_method_for_pulsar_candidate_classification/figures?lo=1&utm_source=google&utm_medium=organic (visited on 06/03/2024).

[21] (), [Online]. Available: https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/ (visited on 06/04/2024).

Curriculum Vitae

FIRST MEMBER

Name-Surname: Beyda GÜLER
Birthdate and Place of Birth: 12.11.2001, Kastamonu
E-mail: beyda.guler@std.yildiz.edu.tr
Phone: +90 553 028 36 36
Practical Training: YTU Probabilistic Robotics Research Group
TUSAŞ A.Ş.
Architecht

SECOND MEMBER

Name-Surname: Şeymanur KORKMAZ
Birthdate and Place of Birth: 04.10.2002, İstanbul
E-mail: seymanur.korkmaz@std.yildiz.edu.tr
Phone: +90 537 497 01 87
Practical Training: YTU Probabilistic Robotics Research Group
ASELSAN A.Ş.
TÜBİTAK BİLGEM

Project System Informations

System and Software: Windows İşletim Sistemi, Python
Required RAM: 2GB
Required Disk: 256MB