escola superior de tecnologia e gestão
instituto politécnico de leiria

Analysing Emotions During Interrogations: A Case Study

Multimodal Sentiment Analysis with Deep Learning

Master in Data Science

JOSÉ ANTÓNIO DA SILVA PINTO GARCIA

Leiria, June 2025

escola superior de tecnologia e gestão
instituto politécnico de leiria

Analysing Emotions During Interrogations: A Case Study

Multimodal Sentiment Analysis with Deep Learning

Master in Data Science

JOSÉ ANTÓNIO DA SILVA PINTO GARCIA

Dissertation conducted under the guidance of Professor Rolando Lúcio Germano Miragaia, Professor Carlos Fernando de Almeida Grilo and Professor Patrício Rodrigues Domingues

Leiria, June 2025

Originality and Copyright

I declare, under oath, that the work presented in this dissertation, with the title “Analysing Emotions During Interrogations: A Case Study”, is original and is made by José António da Silva Pinto Garcia (2232565), under the guidance of Professor Rolando Lúcio Germano Miragaia Ph.D., Professor Carlos Fernando de Almeida Grilo Ph.D. and Professor Patrício Rodrigues Domingues Ph.D..

ACKNOWLEDGMENTS

A special acknowledgment goes to Lieutenant Colonel Tomás from PJM, whose support was instrumental in making the INTU-AI application possible and for providing the necessary resources for its development. I am also grateful to the Criminal Investigation Department of PJM and the director of the 1st Inspector Training Course of 2025, for the time and human resources they dedicated to conducting the functional tests of INTU-AI.

I extend my appreciation to the authors of the publicly available datasets that made this research possible.

Finally, a heartfelt thank you to my academic advisors for their guidance throughout this dissertation and their continuous encouragement during the entire process.

ABSTRACT

In a context of rapid technological advancement and increasing integration of Artificial Intelligence (AI) across various fields, this work explores the application of AI in the domain of criminal investigation, specifically within the context of interrogations. The research presents a dual and complementary nature: the development of a software application for the Portuguese Military Judiciary Police (PJM) and a scientific contribution in the field of multimodal emotion analysis.

The first component focused on the design and implementation of the INTU-AI (Intuition Artificial Intelligence) program, a tool aimed at supporting Military Judicial Police (PJM, from Portuguese Polícia Judiciária Militar) investigators by digitizing and automating administrative procedures related to interrogations. INTU-AI integrates models for Facial Emotion Recognition (FER), Speech Emotion Recognition (SER), and Text-Based Emotion Analysis, functioning as a complete end-to-end solution.

The second component represents a scientific contribution in the form of a proof-of-concept study for dynamic multimodal emotion analysis. Due to the lack of publicly available datasets of criminal interrogations, the MELD (Multimodal EmotionLines Dataset) was employed as the experimental basis, given its resemblance to real-life interaction contexts. This part of the work, structured according to the CRISP-DM methodology, explored three hypotheses regarding the relative importance of each modality in emotional evaluation.

Keywords

Early/Late Fusion, Multimodal Sentiment Analysis, Transfer Learning, INTU-AI, Supervised Learning, Deep Learning, Transformer Models

RESUMO

Num contexto de rápido avanço tecnológico e crescente integração da Inteligência Artificial em diversos domínios, o presente trabalho explora a aplicação da IA na área da investigação criminal, especificamente no contexto de interrogatórios. A investigação assume uma dualidade de naturezas e complementares entre si: o desenvolvimento de uma aplicação informática para a Polícia Judiciária Militar e uma contribuição científica no campo da análise multimodal de emoções.

A primeira componente centrou-se na conceção e implementação do programa INTU-AI (*Intuition Artificial Intelligence*), uma ferramenta destinada a apoiar os investigadores da PJM através da digitalização e automatização de procedimentos administrativos relacionados com interrogatórios. O INTU-AI integra modelos de Reconhecimento Facial de Emoções, Reconhecimento de Emoções na Fala e Análise Emocional Baseada em Texto, funcionando como uma solução completa de ponta a ponta.

A segunda componente constitui uma contribuição científica sob a forma de um estudo de prova de conceito para análise multimodal dinâmica de emoções. Dada a inexistência de conjuntos de dados públicos relativos a interrogatórios criminais, foi utilizado o MELD (*Multimodal EmotionLines Dataset*) como base experimental, devido à sua semelhança com contextos reais de interação. Esta parte do trabalho, estruturada segundo a metodologia CRISP-DM, explorou três hipóteses relacionadas com a importância relativa de cada modalidade na avaliação emocional.

Palavras Chave

Fusão Antecipada/Tardia, Análise Multimodal de Sentimentos, Transferência de Aprendizagem, INTU-AI, Aprendizagem Supervisionada, Aprendizagem Profunda, Modelos Transformer

Index

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

2D	*Two-dimensional*
3D	*Three-dimensional*
Acc	*Accuracy*
AI	*Artificial Intelligence*
ASR	*Automatic speech recognition*
CC	Citizen Card
CNN	*Convolutional Neural Networks*
CREMA-D	*Crowd-sourced Emotional Multimodal Actors Dataset*
CRP	*Convultion Relu Pooling*
DFER	*Dynamic Facial Emotion Recognition*
FER	*Facial Emotion Recognition*
FR	*Facial Recognition*
GLoVe	*Global Vectors for Word Representation*
GRU	*Gated Recurrent Unit*
HOG	*Histogram of Oriented Gradients*
IEMOCAP	*The Interactive Emotional Dyadic Motion Capture*
LFW	*Labeled Faces in the Wild*
LSTM	*Long short-term memory*
MELD	*Multimodal EmotionLines Dataset*
MFCCs	*Mel-Frequency Cepstral Coefficients*
ML	*Machine Learning*
MLP	*MultilayerPerceptron*
MMOD	*Max-Margin Object Detection*
MTCNN	*Multi-task Cascaded Convolutional Networks*
NB	*Naïve Bayes*
NER	*Named Entity Recognition*
NLP	*Natural Language Processing*
PCA	*Principal Component Analysis*
PJM	*Military Judicial Police (Polícia Judiciária Militar)*
POS	*Part-of-Speech*
RAVDESS	*Ryerson Audio-Visual Database of Emotional Speech and Song*
RFB	*Receptive Field Block*

RNN	*Recurrent Neural Networks*
SER	*Speech Emotion Recognition*
SFER	*Static Facial Emotion Recognition*
SSD	*Single Shot Multibox Detector*
SVC	*Support Vector Classifier*
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency - Inverse Document Frequency*
TM	*Text Mining*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
VGG	*Visual Geometry Group*
ViViT	*Video Vision Transformer*
WIDER	*Web Image Dataset for Event Recognition*

1. Introduction

In a world of rapidly evolving technology, where terms like digitalization and AI are continuously redefined and gaining greater significance, companies, organizations, and individuals are becoming voracious consumers of these innovations. These technologies are now capable of surpassing tasks and abilities that were once the sole domain of humans.

Mustafa Suleyman [1] uses the metaphor of a wave in *The Coming Wave,* stating: *"The coming wave is defined by two core technologies: artificial intelligence (AI) and synthetic biology."* This conveys the idea of an unstoppable and continuous force shaping the future. In this context, the evolution and adoption of new technologies become increasingly crucial to keeping pace with global paradigms. Among the various fields where AI has gained prominence, law enforcement and criminal investigations have leveraged these tools, often in areas related to cybersecurity. However, these technologies can and should be applied to broader domains within policing and criminal justice.

This study based on the context of police interrogations, a domain that, despite advances in digital technologies, remains predominantly reliant on human judgment and manual processes. An interrogation typically involves direct verbal interaction between two key actors: the investigator and the suspect.

The procedure encompasses multiple stages, including pre-interrogation preparation, real-time questioning, recording of statements, post-session behavioural analysis, and the drafting of a formal report. Given its legal implications, the process demands precision, attention to detail, and significant time investment. Currently, interrogations are often supported by audio recordings, and in some cases, video footage is also captured when technical conditions permit. However, the responsibility for interpreting suspect behaviour including nuanced facial expressions or emotional cues falls entirely on the investigator, who may sometimes rely on external expert consultation. This traditional approach highlights an opportunity for digital augmentation through intelligent, automated tools.

Emotion recognition is a complex challenge, as human emotions manifest across multiple modalities: facial expressions, Facial Emotion Recognition (FER), voice tone, Speech Emotion Recognition (SER), and word choice (Emotion Analysis). Current deep learning models perform well in each of these areas independently, but there are limited studies

exploring a transfer learning model that integrates all three modalities trained on distinct datasets.

1.1. Objectives

This research presents a solution that contributes to the digital transformation of criminal investigations, specifically in the interrogation process, using the PJM (PJM, from Portuguese *Polícia Judiciária Militar*) as a case study. Two main objectives drove this work:

1. Development of a Software System for Digitalization of Interrogations: The first component focuses on designing and implementing a software solution to digitize administrative procedures related to interrogations. This system utilizes Natural Language Processing (NLP) techniques to streamline documentation while integrating existing models for FER, SER, and emotion analysis from text. By doing so, the system assists investigators in identifying emotions expressed by individuals during interrogations.
2. Scientific Contribution: Given the unavailability of datasets containing criminal interrogation recordings, this component adopts a proof-of-concept methodology. Instead of relying on real interrogation data, the work explores three distinct hypotheses each corresponding to a potential multimodal architecture. These serve as experimental foundations, which could later be applied and refined using real-world interrogation datasets to develop a robust multimodal sentiment analysis model.

1.2. Contributions

The main contributions of this work are multifaceted, encompassing both practical applications and scientific advancements. Firstly, it presents a comprehensive analysis of the state of the art in the field of multimodal models for dynamic emotion recognition, providing a solid theoretical foundation for further research. Secondly, it includes the development of a benchmark model for the AffectAlchemy dataset, contributing a new performance baseline to the community. Another significant outcome is the creation of the INTU-AI (Intuition Artificial Intelligence) software platform, an end-to-end tool designed to support police interrogation processes, which has been made available as open-source software [2]. In parallel, a multimodal approach for emotion analysis was developed and validated through a proof-of-concept study, also released as open-source software [3].

Additionally, the scientific impact of this work is demonstrated through the publication of multiple peer-reviewed papers in international journals and conferences. These include: INTU-AI, a digitalization program for police interrogations; A Comparative Study of Multimodal Emotion Analysis Techniques Using AffectAlchemy; and Design and Implementation of a Multimodal Emotion Analysis Model: A Proof of Concept. Together, these contributions reflect the dual focus of this project addressing a real-world need while advancing the scientific understanding of multimodal emotion analysis.

1.3. Organization of the document

This sequence of chapters is structured to clearly and progressively present the two central pillars of the project: the development of the INTU-AI system and the scientific contribution through multimodal models for emotion recognition and deception detection. Chapter 2 introduces the conceptual background and reviews related work that supports both components. Chapter 3 provides a comprehensive state-of-the-art analysis, highlighting the models used, the latest advances in emotion analysis, and current methodologies in multimodal sentiment and deception detection. Chapter 4 outlines the methodological approach adopted, presenting a practical roadmap for building a real-world end-to-end solution tailored to the operational needs of the Military Judiciary Police (PJM), while also detailing the scientific innovation introduced through the proposed multimodal architecture.

Chapters 5 and 6 delve into the practical implementation and technical innovation at the core of the project. Chapter 5 presents the architecture and functionalities of the INTU-AI system, with a focus on the integration of FER, SER, and text-based emotion models into a unified and customized platform. Chapter 6 shifts towards a more research-driven analysis, introducing multimodal modelling strategies including early and late fusion, transfer learning, and spatially-aware vector representations. Finally, Chapter 7 concludes the thesis by summarizing key findings and suggesting directions for future work.

2. Background

To discuss the developed application INTU-AI[1], it is important to first address the key technological areas it encompasses. Conceptually designed to analyse what the interrogated individual says in terms of content, how they express themselves through facial expressions, and how they convey their speech through tone and delivery, INTU-AI integrates three major fields: Emotion Analysis, facial emotion recognition and speech emotion recognition.

When discussing FER, it is inevitable to touch on related concepts such as facial recognition (FR). Similarly, when exploring Emotion Analysis within the context of spoken interrogations, it becomes necessary to address the technologies that enable speech-to-text conversion.

Since the INTU-AI project encompasses multiple subfields of NLP, we dedicate a specific section to explore these areas in depth. This section will cover both Emotion Analysis from Text, the Summarization Process and the audio-to-text process. This subchapter aims to provide a concise overview of these fundamental concepts.

2.1. Face Recognition

Facial recognition is, in essence, a technology that enables a computer to digitally recognize the image of a person's face. It is not a recent research topic as there are studies back to 1970's [4] in this field. In practice, this technology relies on a specific field of computer science known as computer vision, giving the computer the ability to identify human faces by recognizing their essential features.

The models such as OpenFace, FaceNet, and DeepFace, to name just a few, identify key features of human faces, such as the position of the eyebrows and nose, the openness and positioning of the mouth (including jawline and cheekbone), and even the openness of the eyes [5]. This kind of algorithms are broadly applicable across various fields, including Education, Crime Detection, Healthcare, Public Safety, and others.

Over time, the creation and application of models in various fields have led to exponential growth in this area, resulting in numerous models capable of identifying human faces. Scientific articles suggest that advancements in this field have been so significant that, in

[1] Intuition + IA

some cases, machines can surpass humans in facial recognition [6]. Currently, the process of identifying a human face involves four main stages: i) detection, ii) alignment, iii) representation, and iv) verification [7].

In the first stage, detection, Face Detectors are used to locate faces in images or videos, ignoring everything else. Typically, the detected faces are represented by bounding boxes, or in more advanced models, by the precise cropping of the identified face. There are five tools commonly used for this process: OpenCV[2], MTCNN[3], SSD[4], Dlib[5] and RetinaFace[6].

In the second stage, face alignment becomes relatively simple once the face and eyes are detected. Research has shown that using face alignment can improve model accuracy by over 1% [8]. However, in their original designs, face detectors like OpenCV and Dlib do not offer face alignment as a built-in feature. On the other hand, RetinaFace does and can even detect faces in crowds and accurately locate facial landmarks, including eye coordinates, reason why its alignment performance is notably high compared with others. Facial recognition models nowadays combine all the mentioned techniques.

Deep learning only appears in the representation stage. Instead of classification, the goal is to send face images into a convolutional neural network (CNN) model, to extract embeddings, similar to how autoencoders work.

Verification is the final stage of the process and can be determined through vector comparison or cosine distance. Vector comparison is a simplified process, using Euclidean distance (based on the Pythagorean theorem) to measure the difference between the face representation vectors.

In summary, we can state that face recognition is a combination of CNNs, autoencoders, and transfer learning methodologies.

---

[2] Developed by Paul Viola and Michael Jones in 2001 [90], this algorithm was certainly a gateway for other methods that emerged later, such as HOG + Linear SVM.
[3] Developed by a group of researchers [90], MTCNN was re-implemented using Keras by Iván de Paz Centeno [87].
[4] The SSD was originally developed by a group of researchers from Zoox, Google, and the University of Michigan [90]. In its design, the model does not identify facial landmarks on its own and requires the use of OpenCV resources for this functionality.
[5] Dlib includes two face detection methods built into the library, HOG and MMOD, the first one is for a faster detector, while at the same time being efficient, and the MMOD is for more robust detection.
[6] As part of a project related to ArcFace, some researchers involved in the project developed the RetinaFace [90] model, which was re-implemented using Keras by Stanislas Bertrand [87].

2.2.Facial Emotion Recognition

Facial Emotion Recognition can be defined as a technology used to analyse emotions from sources such as images or videos. This technology is part of affective computing, which according to European Data Protection Supervisor [9] can be described as: "a multidisciplinary field of research on computer's capabilities to recognise and interpret human emotions and affective states, and it often builds on Artificial Intelligence technologies."

In practice, facial expressions are a form of non-verbal communication that convey human emotions and have been the subject of study in two fields of knowledge: psychology and Human-Computer Interaction, more specifically Computer Vision. This topic is not entirely recent, as there are scientific publications addressing it in both Human-Computer Interaction and Psychology, where a clear emphasis on the subject exists: "(…) they offer the prospect of linking two types of elements that are prominent in reactions to emotion, articulate verbal descriptions and explanations and responses that are felt rather than articulated, which it is natural to think of as sub symbolic." [10]

With the proliferation of machine learning, particularly in areas such as deep learning, and alongside the technological development of all electronic devices that have advanced over the past decade, the concept of FER has gained new momentum.

In this context, we can state that the concept of FER currently relies fundamentally on three areas of knowledge: computer vision, machine learning, and psychology, which enable machines to read human facial expressions and classify them according to specific characteristics. The following figure aims to enhance the understanding of the previous concept.

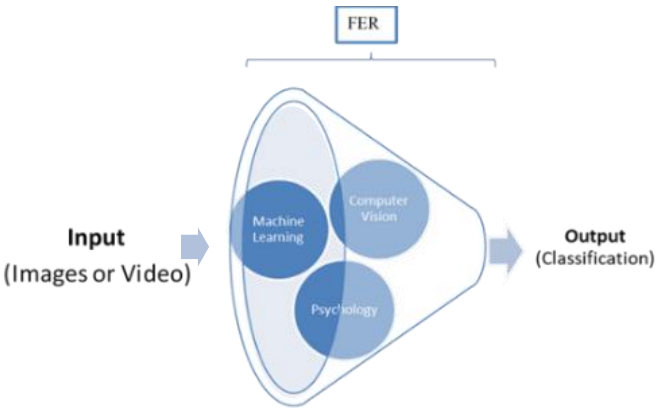


Figure 1 - Facial Emotion Recognition areas

In practice, this FER model allows machines to classify human emotions by analysing facial expressions. To achieve this, as we can see in figure 2 the machine must follow a set of logical steps to provide the desired output, which are: Face Detection, Facial Expression Detection, and Expression Classification.

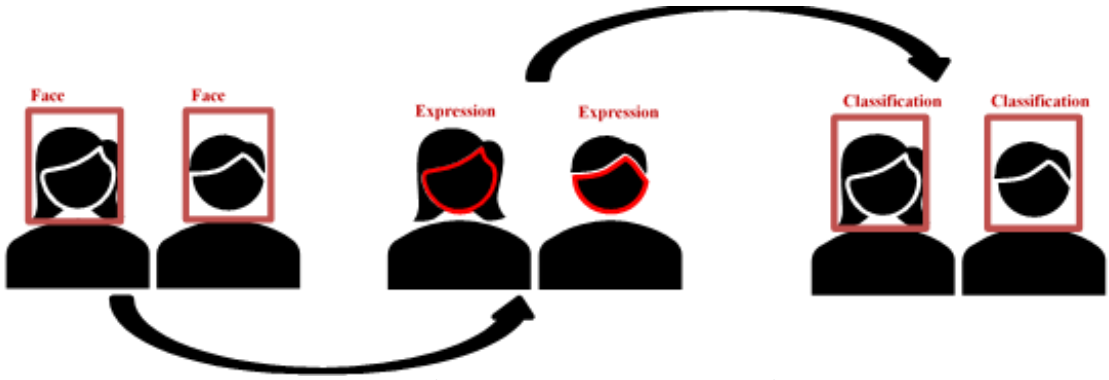


Figure 2 - Steps of Facial Emotion Recognition (adapted from [9])

The proliferation of models focusing on the study of emotions has been a constant over time. In 2009, Bettadapura, V. [11]. included in a survey, a table highlighting 19 of the most relevant works in the field, covering the period from 2001 to 2009. Over the years, more models and model combinations have been developed, leading to the standardization of a three-stage framework for constructing FER systems: Pre-processing, Deep Networks for Feature Learning, and Facial Expression Classification [12], as we can see at the next figure.

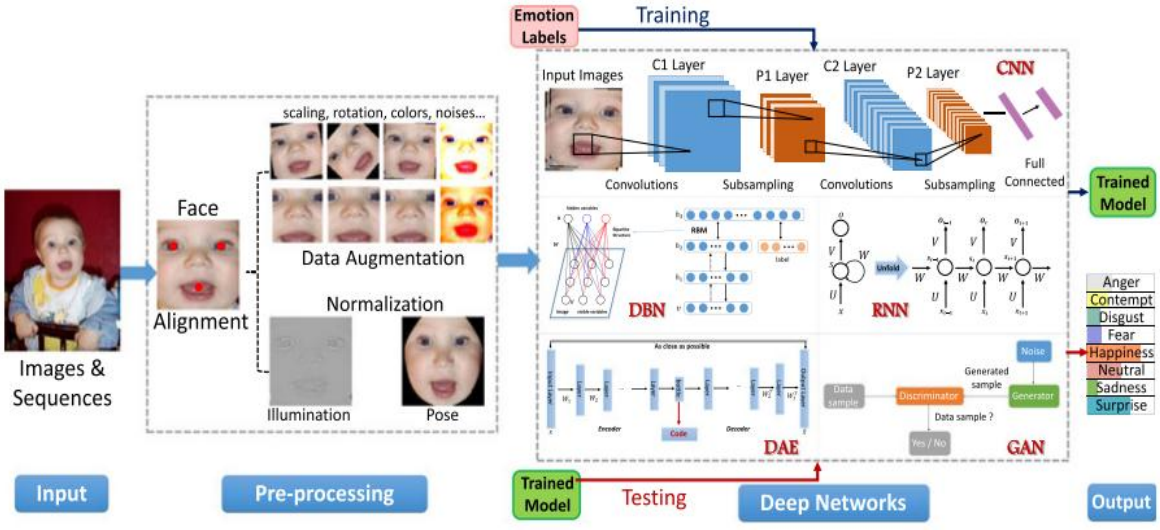


Figure 3 - The general pipeline of deep facial emotions recognition systems (source: [12])

2.2.1. Deep FER Networks for Static Images vs Deep FER Networks for Dynamic Image Sequences

In parallel with the evolution of pipelines, the development of models has also progressed, necessitating the distinction between two subgroups within FER:

- Deep FER Networks for Static Images, which focus on models designed for analysing single, non-sequential images.
- Deep FER Networks for Dynamic Image Sequences, which leverage the temporal correlations between consecutive frames in a sequence, offering significant advantages by incorporating motion cues and temporal patterns into FER systems.

The following table summarizes these two subgroups:

Table 1 - Difference Between Deep FER Networks for Static Images and Deep FER Networks for Dynamic Image Sequences

Subgroup	Description	Key Characteristics
Deep FER Networks for Static Images	Designed for recognizing emotions in isolated static images.	Relies on CNNs; typically applied to datasets like FER2013 or AffectNet[7]; lacks temporal insights.
Deep FER Networks for Dynamic Image Sequences	Focuses on analysing sequences of frames, capturing temporal relationships for enhanced emotion recognition in videos.	Utilizes models such as LSTMs, GRUs, or Transformers; benefits from dynamic context and motion information.

A set of FER models exists for both presented subgroups, employing various techniques and different datasets. There is no doubt that the logical evolution of FER involves the application of Deep FER Networks for Dynamic Image Sequences. The following table summarizes the different types of methods for each.

[7] Affect from the InterNet

Table 2 - Comparison of different types of methods for dynamic and static images adapted from [12]

Deep FER Networks for Static Images							
Network Type		**Data**	**Variations[8]**	**Identity Bias**	**Efficiency**	**Accuracy**	**Difficulty**
Auxiliary Block		Varies	Good	Varies	Varies	Good	Varies
Loss Layer		Fair	Good	Varies	Varies	Good	Varies
Network Ensemble		Low	Good	Fair	Low	Good	Medium
Multitask Network		High	Varies	Good	Fair	Varies	Hard
Cascaded Network		Fair	Good	Fair	Fair	Fair	Medium
GAN		Fair	Good	Good	Fair	Good	Hard
Deep FER Networks for Dynamic Image Sequences							
Network Type		**Data**	**Spatial**	**Temporal**	**Frame length**	**Accuracy**	**efficiency**
Frame Aggregation		Low	Good	No	Depends	Fair	High
Expression intensity		Fair	Good	Low	Fixed	Fair	Varies
Spatio-temporal network	**RNN**	Low	Low	Good	Variable	Low	Fair
	C3D[9]	High	Good	Fair	Fixed	Low	Fair
	FLT[10]	Fair	Fair	Fair	Fixed	Low	High
	CN[11]	High	Good	Good	Variable	Good	Fair
	NE[12]	Low	Good	Good	Fixed	Good	Low

2.3. Speech Emotion Recognition

Speech emotion recognition is a subfield of AI focused on identifying and classifying human emotions from audio signals. The primary goal of SER is to extract acoustic features from speech, such as pitch, rhythm, intensity, and intonation, and map these characteristics to predefined emotional categories.

The ability to understand emotions from speech has broad applications across multiple domains, including healthcare, marketing, security, and human-computer interaction. Traditional classification models such as SVM, Random Forests, and Hidden Markov Model (HMM) were widely employed for mapping these features to emotional labels.

Early studies emphasized the crucial role of Mel-Frequency Cepstral Coefficients (MFCCs) and Bayesian Networks in capturing speech emotion patterns, highlighting the importance of frequency-based representations in emotional state detection.

With the advent of deep learning, models such as CNNs, recurrent neural networks (RNNs), Long Short-Term Memory (LSTMs), and Transformers have come to dominate the field of

[8] Head Pose, Illumination, Occlusion, and Other Environment Factors
[9] 3D ConvNet
[10] Facial Landmark Trajectory
[11] Cascaded Network
[12] Network Ensemble

SER. These approaches enable the automatic extraction of features directly from spectrograms, significantly reducing the need for manual feature engineering [13].

Transfer learning has revolutionized SER by allowing the adaptation of models initially trained on large speech datasets for tasks such as emotion classification. Models like Wav2Vec2, Whisper, and HuBERT, originally developed for automatic speech recognition, have been successfully repurposed for emotion detection tasks [14, 15]. This shift has paralleled the broader evolution of SER, moving away from traditional approaches based on manual feature extraction and classical machine learning toward deep learning and transfer learning paradigms. Transformer-based architectures, particularly Wav2Vec2 and Whisper, have demonstrated strong performance by minimizing the reliance on handcrafted features and achieving high accuracy in emotion recognition. Commonly used databases in SER research include CREMA-D [16], RAVDESS [17] and IEMOCAP [18], which have provided the foundation for the development and benchmarking of modern SER models.

The use of multimodal datasets and the integration of SER with other information sources, such as facial recognition and textual emotion analysis, are further expanding the applications of this technology, enhancing its robustness and real-world usability [19].

2.4. Natural Language Processing

This section covers Emotion Analysis from Text, which focuses on identifying and classifying emotions in verbal content extracted from interrogations, and the summarization process, which is used to automatically generate concise summaries of reports, ensuring that key information is efficiently captured and presented. Additionally, it will also address the Audio-to-Text process, specifically leveraging Whisper to transcribe interrogation audio into text, enabling further analysis through NLP.

2.4.1. Audio-to-Text - Whisper

To analyse the content of spoken text, verbal speech needs to be converted into written text. Several tools are available for this purpose, including both API-based and pre-trained models such as DeepSpeech, SpeechRecognition, Vosk, Wav2Vec2, and Whisper.

According to various studies [20] evaluating the performance metrics of these models, Wav2Vec2 and Whisper consistently stand out as the most accurate and effective solutions.

Due to its ease of use, accessibility, and the availability of high-quality pre-trained models, we have chosen Whisper as our preferred automatic speech recognition (ASR) model for this implementation.

Whisper is an ASR model from OpenAI [21]. It is a robust system with proven reliability, having been tested across numerous hours of transcription in various languages. Its end-to-end architecture, based on an encoder-decoder transformer, processes the input audio by dividing it into 30-second segments and converting them into a spectrogram known as log-Mel spectrogram [22].

According to OpenAI, Whisper was trained on three main elements: the first was multitask training data, with over 680,000 hours of audio featuring English transcriptions, the second transcriptions of spoken English translated into other languages and the last one, multilingual transcription with audio in various languages with transcription in the original language.

The model architecture, as previously mentioned, processes audio by splitting it into 30-second clips and converting it into a log-Mel spectrogram. This spectrogram is then processed by encoders, followed by a decoding step to produce tokens. In summary, Whisper integrates speech recognition, translation, and transcription using an efficient Transformer-based architecture. The whisper model can be resumed in the next figure:

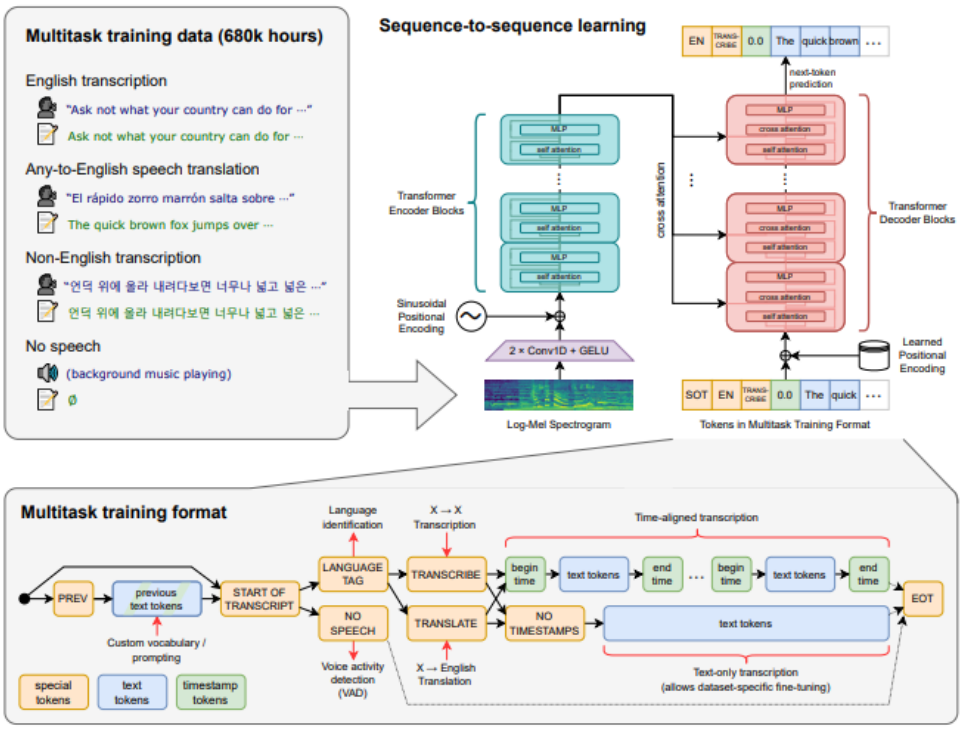


Figure 4 - Summary of the approach defined for the construction of Whisper (source: [18])

2.4.2. Summarization

Summarization is a task aimed at generating a concise and coherent summary from a given text while preserving its most critical information. It is widely used in various applications, such as news aggregation, document summarization, legal and financial reporting, and intelligence analysis. There are two approaches that can be followed:

- Extractive Summarization: This approach selects and directly extracts key sentences or phrases from the original text without modifying them. Examples include classical machine learning models, such as TextRank [23], and deep learning models like BERTSUM [24].
- Abstractive Summarization: The abstractive summarization generates new sentences that convey the essence of the original text using natural language generation. Transformer-based models, such as BART [25] and T5 [26], have demonstrated state-of-the-art performance in abstractive summarization.

One of the models for abstractive summarization in Portuguese is rhaymison/flan-t5-portuguese-small-summarization, a fine-tuned variant of Google's FLAN-T5 adapted specifically for text summarization in Portuguese. This model is based on T5, which reframes all NLP tasks into a unified text-to-text format.

Summarization in NLP has evolved significantly, with Transformer-based models like FLAN-T5 leading the field in abstractive summarization.

2.4.3. Emotion Analysis from Text

Emotion Analysis from Text aims to identify and classify the emotions and sentiments expressed in written or spoken text [27]. This type of analysis is widely used in various applications, such as opinion mining in social media, customer service, and user feedback analysis.

Before proceeding, it is important to clarify the distinction between Sentiment Analysis and Emotion Analysis. Sentiment Analysis classifies a text into categories such as positive, negative, or neutral, based on the emotional polarity expressed through words and phrases. Emotion Analysis goes beyond sentiment analysis by detecting specific emotions such as happiness, sadness, anger, fear, surprise, and disgust.

Emotion analysis is often grounded in psychological models of emotions, such as Plutchik's Wheel of Emotions [28] or Ekman's Model [29] , which identify a set of fundamental emotions.

Early approaches to sentiment analysis relied on lexicon-based methods, using word lists associated with emotions, such as WordNet-Affect [30] and SentiWordNet [31]. These lexicons assign sentiment or emotion scores to words, enabling analysis based on the presence of emotionally charged words in a given text.

However, lexicon-based approaches have significant limitations, particularly in capturing the context in which words appear. For example, words like “great” and “well” can have different meanings depending on the sentence.

The evolution of Emotion Analysis has transitioned from lexicon-based approaches to supervised learning models, which have proven to be more effective. For example, [32] presents models utilizing SVM and Naïve Bayes, while [33] explores n-gram models and logistic regression. These models rely on TF-IDF and bag of words techniques to convert text into a numerical format suitable for classification.

This shift towards machine learning-based methods has allowed for better contextual understanding of emotions in text, addressing some of the limitations of lexicon-based models, such as ambiguity and polysemy.

The advent of deep learning has revolutionized the field of emotions analysis, as RNNs and CNNs have demonstrated superior performance in emotion classification [34]. These innovations eliminate the need for manual feature engineering, allowing models to learn patterns directly from text data.

In recent years, transfer learning and transformer-based models, such as BERT [35] and RoBERTa [36] have achieved state-of-the-art performance in emotion analysis.

Recent studies, such as [37], have demonstrated that fine-tuning BERT and RoBERTa on datasets containing tweets and movie reviews leads to highly efficient models for emotion classification. These models leverage contextual embeddings and transfer learning to accurately capture subtle emotional cues in textual data, outperforming traditional machine learning approaches.

By training on domain-specific datasets, these transformer-based models enhance their ability to understand sentiment nuances, making them particularly effective in applications like social media analysis, customer feedback evaluation, and psychological sentiment detection.

2.5. Available technologies for multimodal emotion analyses from video

Multimodal emotion analysis in video-based systems leverages multiple sources of data, such as facial expressions, vocal tone, and textual sentiment, to enhance emotion detection accuracy. Unlike static image-based approaches, video-based systems must account for temporal dependencies, tracking emotional changes over time. This dynamic aspect introduces additional challenges, such as handling subtle micro expressions, speech prosody, and contextual variations [38]

Recent advancements in deep learning, RNNs, and transformer architectures have significantly improved the ability to analyse emotions in videos. The integration of FER, SER, and text-based emotions analysis has paved the way for sophisticated real-time emotion recognition systems. This section explores cutting-edge technologies specifically designed for video-based multimodal emotion analysis.

2.5.1. Facial Emotion Recognition in Video

Video-based FER requires tracking and analysing facial expressions over time to capture emotional transitions and patterns. Unlike static, image-based FER, which classifies emotions from single frames, video-based FER integrates spatio-temporal analysis, ensuring that transient facial distortions, such as a brief smile or a raised eyebrow, are not misclassified.

One prominent approach for video-based FER is the use of 3D-CNNs, which extend traditional 2D convolutional networks by incorporating temporal convolutions, allowing the model to extract spatio-temporal features from video sequences. A notable example is ResNet3D-18, an adaptation of the ResNet architecture designed specifically for FER tasks, where convolutional layers learn both spatial and temporal facial expression patterns [39].

Another widely adopted approach combines CNNs for spatial feature extraction with RNNs for temporal analysis. In this method, CNNs (e.g., VGG-Face, ResNet-50) are used to extract

spatial features from each frame, while RNNs, such as LSTMs or gated recurrent units (GRUs), model temporal dependencies in facial expressions.

An example of this architecture is illustrated in Figure 5 [39], where the system integrates two GRU layers on top of CNN-extracted features, enhancing the network's ability to capture emotional transitions over time [40].

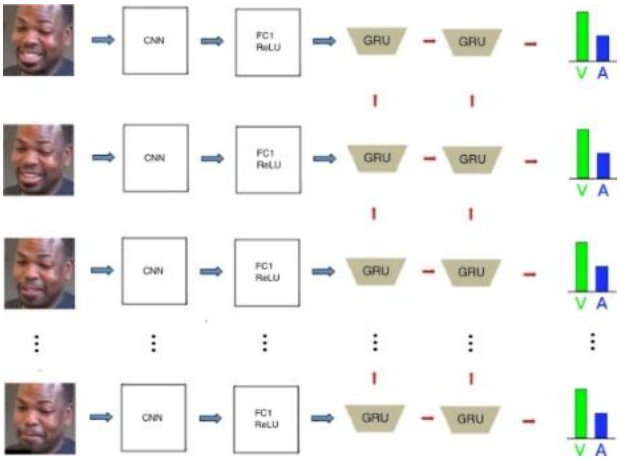


Figure 5 - CNN + RNN approach [39]. (V- valence , A- arousal)

A more recent and advanced approach is transformer-based video FER models, which utilize self-attention mechanisms to process entire video clips and capture long-range dependencies. One such model is ViViT, which applies spatial-temporal attention to extract emotions from video sequences holistically [41]. Another significant model, TimeSformer, eliminates recurrence and instead processes video frames using a purely transformer-based spatial-temporal framework, enabling efficient and robust emotion classification [42].

To ensure accurate FER, the first step is detecting and tracking faces across video frames. Techniques such as MTCNN are commonly employed to detect and normalize faces for further analysis. Another approach is OpenFace, developed by researchers at Carnegie Mellon University [43], which dynamically tracks facial landmarks, ensuring robust detection even under varying pose, lighting, and occlusion conditions.

Once faces are detected, extracting facial features across multiple frames is crucial. The constrained local neural fields (CLNF) model, integrated into OpenFace, is widely used to extract action units, fine-grained facial muscle movements that define different emotions. These extracted features serve as the foundation for emotion classification in video-based FER.

To enhance recognition accuracy, temporal emotion aggregation techniques are applied. One effective approach involves temporal attention mechanisms, which prioritize key facial expressions within the video, ensuring that the most relevant frames contribute more significantly to the final emotion prediction. Additionally, GNNs are leveraged to model relationships between facial landmarks across time, capturing the dynamic progression of facial expressions.

By combining these key techniques, video-based FER systems achieve improved accuracy in real-time emotion recognition, making them suitable for applications such as forensic investigations, psychological assessments, and human-computer interaction.

Several publicly available datasets are used to train and evaluate video-based FER models:

- AffectNet (Affect from the InterNet) videos is a large-scale dataset designed specifically for video-based FER, containing annotated emotional expressions across diverse real-world and controlled settings.
- AFEW (Acted Facial Expressions in the Wild) is a dataset that provides video sequences labelled with emotional states, offering a more dynamic and realistic representation of facial expressions [44]
- The Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) is a multimodal dataset including video, audio, and transcriptions, enabling the integration of multiple modalities for more comprehensive emotion recognition [16].
- The Multimodal EmotionLines Dataset (MELD) is an enhanced version of the EmotionLines dataset, incorporating text, audio, and visual modalities. It contains over 1400 dialogues and 13000 utterances from the friends TV series, featuring multiple speakers per dialogue [45].

These datasets play a crucial role in advancing FER research, offering well-annotated emotional expressions captured in various environments, ranging from controlled lab conditions to real-world, spontaneous interactions.

2.5.2. Speech Emotion Recognition in Video

Speech Emotion Recognition (SER) in video-based systems differs from traditional SER by requiring synchronization with facial expressions and text transcriptions. Variations in speech rhythm, intensity, and prosody provide critical cues for emotion detection, making it

essential to integrate multiple modalities for accurate classification. Several deep learning approaches have been developed for video-based SER, each leveraging different architectures to enhance emotion recognition, including self-supervised speech representation learning, hybrid CNN-LSTM models, and transformer-based approaches.

Self-supervised learning models such as Wav2Vec 2.0 [46] and HuBERT [47] have demonstrated strong capabilities in capturing emotional variations in speech without requiring large amounts of labelled data. These models effectively extract meaningful speech representations, which are crucial for emotion classification. Another effective approach involves hybrid CNN-LSTM models, where CNNs extract frequency features from spectrograms, transforming raw audio into a structured format suitable for deep learning, while LSTMs capture temporal dependencies in emotional speech patterns. This combination allows the system to analyse variations in tone, rhythm, and pitch over time, improving the detection of subtle emotional cues.

Transformer-based models have also gained prominence in SER. SpeechT5 [48] unifies multiple speech-related tasks, including emotion recognition, by leveraging self-supervised learning to enhance speech representation across different domains. Similar, WavLM [49] specializes in capturing fine-grained emotion variations in conversational settings, making it particularly useful for real-world applications where subtle shifts in speech tone can indicate emotional changes.

To ensure accurate emotion classification in video-based SER, several key techniques are employed, including speech pre-processing, temporal speech analysis, and multimodal audio-visual synchronization. In speech pre-processing, voice activity detection is used to identify speech segments within the video, eliminating background noise and ensuring that only relevant speech is analysed.

Additionally, MFCCs are extracted to provide spectral audio features, which are crucial for distinguishing between different emotional tones. For temporal speech analysis, a sliding window approach [50] is commonly used to segment audio into 5-10 second intervals, allowing for frame-wise emotion classification and ensuring that emotion recognition remains granular and responsive to real-time speech variations.

Multimodal synchronization techniques further improve emotion recognition by aligning facial expressions with speech cues. Dynamic Time Warping [51], for instance, helps

synchronize speech and facial expressions, ensuring that detected emotions across modalities are temporally consistent. This is particularly important in forensic applications and psychological assessments, where accurate emotion tracking can provide valuable insights into a subject's psychological state.

Several widely used datasets provide training and evaluation benchmarks for video-based SER models:

- The Interactive Emotional Dyadic Motion Capture (IEMOCAP [18]) is a multimodal dataset containing audio-visual emotional expressions, commonly used in emotion recognition research.
- CREMA-D [16]: As explained at 2.2.1 can be used for SER models.
- Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS [17]) includes both facial and vocal emotional expressions, recorded in controlled conditions to ensure high-quality emotional labelling.
- MELD as explained earlier, this dataset can be used for SER models.

These datasets play a crucial role in the development of SER models, providing diverse emotional expressions, variations in speech tone, and multimodal cues that enhance the robustness of video-based emotion recognition systems. By integrating self-supervised learning, hybrid deep learning architectures, and multimodal synchronization techniques, modern SER models achieve improved accuracy and adaptability, making them valuable tools in forensic investigations, human-computer interaction, and psychological research.

2.5.3. Text-Based Emotion Analysis in video

Text-based emotion recognition in video settings involves analysing spoken transcriptions using NLP techniques. When applied in a multimodal context, the objective is to detect the emotional tone embedded in speech and align it with FER and SER outputs to achieve a comprehensive emotional understanding of the subject.

Several state-of-the-art models are widely used for emotion classification through text analysis. RoBERTa [36] is a robustly trained version of BERT, which enhances sentiment analysis performance by leveraging improved pretraining strategies and handling linguistic nuances more effectively. Another commonly used model is DistilBERT [52] a lightweight

alternative to BERT, optimized for real-time applications where computational efficiency is critical.

Additionally, advanced generative models such as [53] and GPT 4 [54] developed by OpenAI, have demonstrated strong capabilities in emotion classification and text-based emotion analysis.

Key techniques used in text-based emotion analysis include Speech-to-Text Transcription and Aligning Text Sentiment with Video-based FER and SER.

Regarding available datasets for text-based emotion recognition in video settings, two stand out as the most widely used for multimodal applications. MELD [45] and CMU-MOSEI [55]. CMU-MOSEI contains videos with aligned textual, vocal, and facial emotion labels, making it highly suitable for training models that integrate multiple modalities for emotion analysis.

2.5.4. Multimodal Fusion for Emotion Recognition in Video

A key challenge in multimodal video-based emotion analysis lies in effectively integrating FER, SER, and text-based sentiment analysis to achieve a comprehensive understanding of emotional expressions. To address this, three primary fusion strategies are commonly employed [56]:

- **Early Fusion**: Features from all modalities (facial, speech, and text) are concatenated at the input level before being processed by a unified model. This approach captures cross-modal dependencies early but may struggle with feature alignment.
- **Late Fusion**: Each modality is processed separately using specialized models, and the results are combined at the decision level. While this method maintains interpretability, it may lose some interactions between modalities.
- **Hybrid Fusion**: A combination of early and late fusion techniques, aiming to balance feature representation and interpretability. This approach allows models to learn both independent and interdependent emotional cues across modalities.

In addition to widely referenced datasets such as MELD and CMU-MOSEI, new datasets continue to emerge to further improve multimodal emotion recognition in video-based systems. Notable examples include OV-MER [57] and MER2024 [58] , which provide richer

and more diverse emotional annotations, facilitating the development of more robust models capable of handling real-world variability.

Multimodal emotion recognition in video-based systems presents unique challenges, requiring a combination of spatio-temporal analysis, multimodal synchronization, and advanced fusion techniques to accurately classify emotions. Advances in deep learning, particularly the use of transformers and self-supervised learning, have significantly enhanced FER, SER, and text-based sentiment analysis, paving the way for more reliable and context-aware emotion recognition models.

2.6. Summary

This background chapter aims to contextualize the fundamental technologies that underpin emotion analysis, with a particular focus on the approaches developed within the INTU-AI project, which constitutes the first phase of this work. The program integrates three key areas: FER, SER, and Text-Based Emotion Analysis, each with its own set of underlying technologies and concepts.

Face recognition is a foundational technology for FER, enabling computers to digitally recognize and process human faces using computer vision techniques. The facial recognition process typically involves four main stages: i) detection, where faces are located in images or videos; ii) alignment, which improves the accuracy of facial recognition models by normalizing detected faces; iii) representation, where feature embeddings are extracted using deep learning models such as CNNs and iv) verification, which compares vectorized face representations for identity recognition.

In contrast, FER focuses on analysing emotions from facial expressions in images or videos, integrating computer vision, machine learning, and psychology to classify facial expressions into specific emotions. FER systems typically follow a three-stage pipeline: pre-processing, which enhances facial image quality and extracts key facial landmarks; feature learning via deep neural networks, where discriminative features are extracted for emotion classification; and facial expression classification, mapping extracted features to predefined emotional categories.

There is a difference between static-image FER models and dynamic-image sequence FER models, with the latter incorporating temporal dependencies, providing a richer analysis by tracking emotional transitions over time. Advanced deep learning models such as 3D-CNNs,

hybrid CNN-RNN architectures, and Transformer-based approaches like ViViT and TimeSformer have significantly improved the accuracy of video-based FER.

Speech emotion recognition aims to extract acoustic features such as tone, rhythm, and intensity from speech and map them to predefined emotional categories. Initially, traditional machine learning classifiers such as SVM and Random Forests were commonly used, leveraging handcrafted features like MFCCs and Bayesian Networks for emotion classification. With the rise of deep learning, CNNs, RNNs, LSTMs, and Transformers have become dominant in SER, allowing for automatic feature extraction directly from spectrogram representations of speech signals. Recent advancements in self-supervised learning, such as Wav2Vec2 and Whisper, have revolutionized SER by leveraging large-scale unsupervised pretraining, leading to more robust emotion recognition even in noisy environments.

NLP plays a crucial role in analysing the verbal content of interrogations. In the context of emotion analysis, NLP is applied across three key tasks: i) speech-to-text transcription, which converts audio into text for further analysis using tools like Whisper known for its high transcription accuracy; ii) text summarization, which generates concise summaries from large text data using extractive and abstractive approaches, with Transformer-based models such as FLAN-T5 demonstrating state-of-the-art performance; and iii) text-based Emotion Analysis, which identifies and classifies emotionally charged language within text.

Emotion analysis detects specific emotions such as anger, happiness, fear and others. The field has evolved from lexicon-based methods to machine learning-based classification models, and more recently, to deep learning and transfer learning techniques such as BERT and RoBERTa, which capture contextual emotional nuances.

By incorporating temporal dependencies, multimodal approaches provide deeper insights into emotional states. Video-based FER analyses facial expressions over time, employing deep learning models such as 3D-CNNs, which extract spatiotemporal features from video frames, and hybrid CNN-RNN models, which combine spatial feature extraction (CNNs) with temporal modelling (RNNs/LSTMs).

Transformer-based approaches, such as ViViT and TimeSformer, capture long-range dependencies in emotion transitions. For effective face detection and tracking in video,

techniques such as MTCNN and OpenFace are employed to ensure consistent and accurate tracking of facial landmarks under varying lighting and pose conditions.

Video-based SER synchronizes speech analysis with facial expressions and text transcriptions to achieve a multimodal understanding of emotional states. Key approaches include self-supervised learning models such as Wav2Vec2 and HuBERT for unsupervised speech representation learning, hybrid CNN-LSTM models that extract spectrogram features and model temporal dependencies in emotional speech patterns, and Transformer-based SER models, such as SpeechT5 and WavLM, which specialize in detecting fine-grained emotional variations in conversational settings.

Speech pre-processing techniques, such as Voice Activity Detection (VAD) and MFCCs feature extraction, play a critical role in enhancing emotion classification. Furthermore, temporal analysis techniques like the Sliding Window Approach allow for frame-wise speech emotion recognition over sequences of five to ten seconds.

Text-based emotion recognition in video settings involves analysing spoken transcriptions using NLP models such as RoBERTa and DistilBERT. This approach enables a deeper understanding of verbal sentiment, aligning it with FER and SER outputs to ensure a coherent multimodal interpretation of emotions.

To integrate FER, SER, and text-based emotion analysis, multimodal systems employ fusion techniques, which include Early Fusion, where raw features from all modalities are concatenated at the input level; Late Fusion, where each modality is processed independently and combined at the decision level; and Hybrid Fusion, which balances feature representation and interpretability.

The integration of these diverse FER, SER, and NLP technologies allows for a comprehensive emotion analysis framework, which is crucial for enhancing the INTU-AI project in the context of interrogations.

The continuous advancements in multimodal emotion recognition, driven by richer datasets and more sophisticated deep learning models, reinforce the growing importance and applicability of this field in forensic analysis, psychological research, and human-computer interaction.

3. Related Work

This chapter reviews the related work that underpins the development of the INTU-AI system. It provides an overview of the key models and datasets employed for each of the three main modalities explored in this project: FER, SER, and Textual Emotion Analysis. It also discusses the transition from unimodal emotion classification to a multimodal emotion recognition approach, highlighting recent advances in the integration of video, audio, and text information for emotion and deception detection. The models and datasets selected for each modality were chosen based on their relevance, performance, and availability, ensuring a solid foundation for building a functional and testable prototype.

3.1. Models used for the development of the INTU-AI program

The primary goal behind the INTU-AI program was to deliver a functional prototype as quickly as possible to the PJM, enabling real-world testing and validation throughout the course of this research. Given this requirement, it was imperative to leverage publicly available pretrained models for the classification of the three core modalities: FER, SER, Emotion Analysis from Text.

Regarding Emotion Analysis from Text, since our initial goal was to analyse emotions exclusively in Portuguese-language texts, and given that datasets related to interrogations are not commonly available in public repositories, we decided to develop a custom model from scratch. A detailed discussion of this model and its implementation will be provided in the dedicated chapter on the PJM application.

3.1.1. Facial Emotion Recognition

For our work, we adapted the FER system available in [59], leveraging two specialized technologies for emotion classification: Face Detection RFB-320, trained on the WIDER FACE dataset for face detection, and a VGG13 model, trained on the FER+ dataset for Facial Emotion Recognition.

Facial Emotion Recognition plus Dataset

The FER+ is an improved version of the FER dataset, originally introduced for the task of automatic facial expression recognition in images. The primary motivation behind the creation of FER+ was to overcome the limitations of the original FER dataset, providing

more refined labelling and a better representation of the complexity of human expressions. This dataset falls within the category of datasets that feed SFER models. The FER+ dataset introduces two new emotion categories: “Neutral” and “Contempt”, in addition to the existing categories from the original FER dataset: Happiness, Sadness, Anger, Surprise, Fear, and Disgust. This modification significantly enhances the dataset, making it more comprehensive and effective for facial expression classification. The manual annotation process, reduced classification errors and increased label reliability, ensuring a more accurate dataset for training emotion recognition models.

The FER+ dataset has been widely used to train CNNs for the task of FER. Models trained on FER+ demonstrate superior performance compared to those trained on the original FER dataset. This dataset represents a significant advancement in the field of static facial expression recognition, providing a more robust and refined dataset for training deep learning models.

Visual Geometry Group - Face

The Visual Geometry Group - Face (VGG-Face) is constructed based on VGGNet framework, which has significantly influenced the field of computer vision since its inception. Introduced by the Visual Geometry Group at the University of Oxford, VGG models gained prominence in the 2014 ImageNet Large Scale Visual Recognition Challenge for their deep CNNs with uniform architecture. VGG-19, the deepest variant, stood out for its simplicity and effectiveness [60].

This structure enables efficient extraction of features from images, making it ideal for facial recognition applications. The VGG Face Model streamlines the original VGG-16 architecture by keeping only three main blocks and removing the final two convolutional blocks. This adjustment reduces the model's complexity while preserving its effectiveness in various facial recognition scenarios [61].

To accomplish this, they provided a new face dataset containing 2.6 million faces with minimal manual labelling. The structure of the VGG-Face model is demonstrated in the next figure.

Figure 6 - Architecture VGG-Face model (adapted [90])

Web Image Dataset for Event Recognition Face Dataset

The WIDER FACE dataset is a large-scale benchmark dataset for face detection, extensively used for evaluating deep learning models in challenging real-world scenarios. It was developed to overcome the limitations of previous face detection datasets by introducing a highly diverse collection of images with extensive variations in scale, pose, occlusion, and facial expressions. It comprises 32203 images, covering a broad spectrum of conditions, including different lighting environments, complex backgrounds, and various degrees of facial occlusion. It is categorized into three difficulty levels: easy, medium, and hard, each representing increasing levels of complexity in face detection due to variations in occlusion, pose, scale, and background clutter.

Over the years, the WIDER FACE dataset has become a gold standard for benchmarking face detection algorithms. It is widely used for training and evaluating models in real-world applications, such as face recognition, surveillance, and biometric authentication.

Face Detection, Receptive Field Block - 320

The RFB-320 is a lightweight yet highly efficient face detection model designed for real-time applications, particularly in edge computing environments. It introduces a modified RFB module, which enhances multiscale contextual information capture while maintaining low computational cost [62]. The model is available in two versions: Version-Slim, prioritizing speed, and Version-RFB, offering higher accuracy. Trained on the WIDER FACE dataset, it achieves state-of-the-art performance across different difficulty levels, with detection accuracies of 78.7% (easy), 69.8% (medium), and 43.8% (hard) at 320×240 resolution [62]. The model supports ONNX, NCNN, MNN, and Caffe, enabling seamless deployment across platforms. The RFB-320 SSD model is widely used in applications

requiring fast and reliable face detection, particularly in resource-constrained environments, making it a strong candidate for real-time FER systems.

3.1.2. Speech Emotion Recognition

For the analysis of SER, we employed a fine-tuned version [63] of the XLSR-53 large model [64], originally designed for English speech recognition. This model was further trained on the RAVDESS dataset, which includes the set of emotional states that align closely with the objectives of our study. The choice of this model was motivated by its proven effectiveness in emotion classification tasks and its compatibility with the linguistic characteristics and emotional nuances present in the dataset.

Ryerson Audio-Visual Database of Emotional Speech and Song dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [17] is a comprehensive multimodal dataset designed to support research in emotion recognition across disciplines such as neuroscience, psychology, and computer science. Unlike most existing datasets which focus on unimodal vocal recordings, RAVDESS captures the multimodal nature of emotional communication. It consists of 7356 clips across three modalities[13] performed by 24 professional actors. It includes eight emotion categories: happiness, sadness, anger, fear, surprise, disgust, calmness, and neutral, each expressed at two intensity levels (normal and strong). The inclusion of both neutral and calm expressions allows for more nuanced experimental conditions.

All recordings were captured in a professional studio using two lexically neutral phrases spoken and sung by each actor, enabling direct comparisons between speech and song. The validation study involved 247 North American participants who evaluated a subset of the dataset for emotion category, intensity, and genuineness. Results demonstrated high emotional validity, with mean recognition accuracies of 0.72 for speech and 0.71 for song. Audio-visual stimuli yielded significantly higher recognition accuracy, and higher intensity expressions were more accurately identified.

In summary, RAVDESS is a validated, publicly available, multimodal emotional dataset in English, uniquely including matched speech and song expressions. Its multimodal nature,

[13] audio-visual, video-only and audio-only.

emotional intensity variation, and high reliability make it a valuable resource for research on emotion perception, processing, and recognition.

Wav2vec2 Large Xlsr 53

The Fine-tuned XLSR-53 model [63], is a specialized adaptation of Facebook's wav2vec2-large-xlsr-53 architecture, fine-tuned specifically for English automatic speech recognition. Designed for audio inputs sampled at 16 kHz, the model performs direct speech-to-text inference using a Transformer-based Wav2Vec2ForCTC architecture. It demonstrates high accuracy in transcribing spoken English from various sources such as podcasts and audiobooks, and it is efficient in handling both short utterances and large-scale audio datasets. The model ranks highly in tasks related to speech recognition, spoken language understanding, and speech enhancement. Despite its strong performance, the model requires input audio to be resampled to 16 kHz and may struggle with speech patterns or accents not sufficiently represented in the Common Voice 6.1 dataset. Additionally, its use is limited to English-language tasks.

Overall, the Fine-tuned XLSR-53 is a state-of-the-art ASR model for English, balancing speed, accuracy, and ease of deployment. Its open availability and compatibility with libraries such as HuggingFace Transformers and HuggingSound make it a valuable resource for speech processing tasks in academic and applied research settings.

3.1.3. Emotion Analysis

To analyse the emotions, present in the text extracted from interrogations, we require a model capable of identifying and classifying those emotions.

While several public datasets are available for building such a model, we chose to train ours using the AffectAlchemy dataset [65] due to its recent publication and strong academic relevance.

In addition to employing well-established classification techniques such as SVC, Naïve Bayes, Random Forest, Decision Tree, XGBoost, KNN, and Logistic Regression, we also explored two deep learning approaches: a CNN with pre-trained GloVe[14] embeddings, and a RNN also using GloVe embeddings. Moreover, we further enhanced our analysis by fine-

[14] GLoVe for English and Fasttext [87] for Portuguese

tuning state-of-the-art pre-trained transformer models, namely DistilBERT [52] and RoBERTa [36], using our dataset.

AffectAlchemy dataset

In 2024, a new affective dataset named "AffectAlchemy" was introduced, designed for emotion recognition in text [65]. The dataset is based on Robert Plutchik's Wheel of Emotions [66] and was subsequently analysed using machine learning techniques. The study falls within the field of affective computing, which leverages artificial intelligence to identify, interpret, and respond to human emotions. Text-based affective systems, a subset of this area, aim to classify textual data into a broader range of emotional categories, such as fear, sadness, anger, and happiness. To achieve this, these systems draw on various psychological models of emotion recognition. The main contribution of this work is the creation and validation of the AffectAlchemy dataset, grounded in Plutchik's theory, which posits eight primary emotions that combine to form more complex secondary emotions, resulting in a total of 32 distinct emotional states. The dataset was built through a semi-automated methodology, combining existing datasets based on Ekman's model, social media text, manually generated text, and literary excerpts. All data underwent cleaning and pre-processing using multiple NLP tools, including NLTK, NeatText, SpaCy, and TextBlob. Emotion labelling was performed with the assistance of LLMs.

In conclusion, the AffectAlchemy dataset aims to contribute significantly to the field of affective computing by offering a valuable resource for the development of models capable of deeper emotion recognition and a more nuanced understanding of emotional context. Its public release on GitHub [67] seeks to support ongoing research in affective systems and the development of emotionally-aware applications.

Regarding the dataset made available by the authors, everything suggests that the version published on GitHub is incomplete. In the paper, the authors mention a dataset containing 31066 instances, whereas the available version includes only 20083. This discrepancy may help explain the variation in performance metrics between the original study and the replication attempt. While the authors report an accuracy of 81.98% and an F1-score of 81.84%, our replication achieved only 66.18% accuracy and 66.55% F1-score, as demonstrated in Section "*Modelo Sentiment analises 4 emocoes.ipynb*".

Convolutional Neural Networks/Recurrent Neural Networks with GloVe pre-training

The CNNs, originally designed for computer vision tasks, have also shown promising results in NLP, particularly in text classification and emotion recognition tasks. According to Dive into Deep Learning [68], the main advantage of CNNs in this context lies in their ability to capture local patterns and n-grams through convolutional filters applied over sequences of word embeddings.

In this approach, each word in the text is represented as a dense vector obtained from a pre-trained model such as GloVe, which preserves both semantic and syntactic relationships between words. These embeddings are typically loaded and used as either a fixed or trainable input layer. The convolutional filters extract relevant features from the sequences, and after pooling operations and dense layers, the model becomes capable of performing emotion classification.

RNNs, on the other hand, are specifically designed to handle sequential data and are particularly effective at modelling the temporal dependencies between words in a sentence. In the context of emotion recognition in text, RNNs are frequently used to capture the dynamic and sequential structure of language. In this approach, GloVe embeddings are also used to initialize the vector representation of each word. Unlike CNNs, RNNs process the text word by word, maintaining an internal state that is updated at each step. This allows the model to capture long-range relationships between terms, which is crucial for understanding emotional nuances that unfold throughout a sentence. Variants such as LSTM and GRU are often preferred over traditional RNNs due to their ability to mitigate the vanishing gradient problem, thereby enabling the learning of longer-term dependencies. The integration of GloVe embeddings provides the model with a rich semantic starting point, which helps accelerate convergence and improve performance, especially in scenarios with limited data.

Both approaches, CNN or RNN both with pre-train GloVe, have their merits and distinct applications in the field of textual emotion recognition. CNNs are generally faster and more effective at detecting local patterns, while RNNs (particularly LSTM and GRU) provide a deeper analysis of the sequential structure of the text.

3.2. Multimodal fusion to Emotion analysis from video

The multimodal integration proposed in the present work goes beyond the isolated classification of the three key modalities capable of identifying human emotions during an interrogation SER, FER, and textual emotion analysis referred to here as Multimodal Emotion Recognition. In addition to emotion classification, this integration aims to support the detection of potential deception or truthfulness, leveraging emotional cues expressed by the subjects under interrogation.

Due to the scarcity of publicly available datasets encompassing all three modalities in this specific context, we adopted a hybrid approach to develop a model capable of identifying not only the underlying emotions but also the veracity of the subject's statements. While multimodal emotion recognition is conceptually well-accepted and there is a growing body of literature exploring this approach from different perspectives, most existing works are still in an experimental phase, especially regarding the joint modelling of SER, FER, and textual sentiment. Nevertheless, several recent studies have achieved promising performance metrics.

In our case, to ensure that all three modalities were represented, we selected the MELD dataset as the foundation for our multimodal approach. As a benchmark for this type of integration, we identified two relevant studies: the first, which explores only SER and textual sentiment analysis, thus not fully aligned with our objectives, and a second study [69] , which specifically adopts a multimodal approach incorporating all three emotion recognition modalities. This latter work was used as a reference for our experimental setup.

3.2.1. Multimodal EmotionLines Dataset

The Multimodal EmotionLines Dataset (MELD) work paper [45] addresses the increasing importance of emotion recognition in conversations highlighting a significant gap in the availability of large-scale multimodal datasets involving more than two participants per dialogue. To address this limitation, the authors introduce the MELD, which extends and improves upon the original EmotionLines dataset.

MELD comprises approximately 13000 utterances from 1433 dialogues sourced from the television series Friends. Each utterance is annotated with both emotion and sentiment labels and includes textual, audio, and visual modalities. The dataset was constructed by extracting

precise timestamps from the original EmotionLines corpus, applying constraints to ensure dialogue coherence, same episode and scene. The utterances were then re-annotated by three human ratters who viewed the video clips, and final labels were determined via majority voting.

The paper provides a detailed breakdown of emotion and sentiment distribution, noting the non-uniform distribution with a predominance of neutral emotions. Key statistics are also presented, including average utterance length, the number of emotions per dialogue, and the frequency of emotion shifts between speakers.

To establish robust baselines, the authors evaluated several models, including: Text-CNN (non-contextual baseline), bcLSTM (unimodal and bimodal contextual model), and DialogueRNN (tracks the emotional state of each speaker throughout the conversation). Among these, the multimodal DialogueRNN achieved the best performance, with a weighted F1-score of 67.56% for sentiment classification and 60.25% for emotion classification. The results emphasize the critical role of context and multimodality in accurately recognizing emotions in conversations. Further analysis reveals the importance of inter-speaker influence, with the DialogueRNN model frequently attending to utterances from other participants in successful predictions. In summary, MELD represents a valuable benchmark for advancing research in multimodal. The strong baselines and insights provided pave the way for more sophisticated and human-like emotion understanding.

3.2.2. Multimodal Emotion recognition in conversations

The work presented by the authors in [69] introduces a novel approach named GraphSmile, designed to address the challenging tasks of Multimodal Emotion Recognition in Conversations (MERC) and Multimodal Sentiment Analysis in Conversations (MSAC). GraphSmile is proposed as a unified framework capable of handling both tasks simultaneously by integrating two core components: Graph Structure Fusion (GSF) and Sentiment Dynamics Perception (SDP).

The GSF module is designed to leverage graph-based structures to alternately assimilate inter-modal and intra-modal emotional dependencies in a layer-wise manner. This architecture enables the model to effectively capture cross-modal cues while mitigating fusion conflicts. In constructing the multimodal dialogue graphs, GraphSmile establishes connections not only between nodes of different modalities within the same utterance, but

also directly across utterances. The GSF employs a simplified graph convolution operation to propagate emotional signals both within and across modalities. Furthermore, the residual connections in GSF help alleviate the over-smoothing problem and allow for the aggregation of multi-hop emotional cues.

On the other hand, the SDP module serves as an auxiliary mechanism that explicitly models sentiment dynamics across utterances. It enhances the model's ability to detect and differentiate abrupt sentiment shifts by adopting a contrastive learning strategy. Specifically, SDP brings together utterances with similar sentiment while distancing those with opposing sentiments. This auxiliary task operates during training and complements the main classification objectives.

GraphSmile is structured as a multi-task learning model, jointly optimizing three loss functions: emotion classification, sentiment classification, and sentiment-shift detection (via SDP). This joint optimization facilitates knowledge transfer between MERC and MSAC, thereby improving the model's sensitivity to emotional and sentimental nuances within multimodal conversations.

To validate the performance of GraphSmile, the authors conducted extensive experiments on three benchmark datasets for multimodal affective computing: IEMOCAP, MELD, and CMU-MOSEI. The results demonstrated that GraphSmile significantly outperforms existing state-of-the-art models in terms of accuracy and weighted F1-score. Moreover, detailed performance analyses revealed that GraphSmile achieves more balanced results across various emotion categories, further supporting its robustness and generalizability.

In summary, GraphSmile introduces an innovative architecture that addresses key challenges in multimodal emotion and sentiment recognition by combining inter- and intra-modal graph fusion with a specialized module for modelling sentiment dynamics. The strong empirical results validate its effectiveness and its contribution to advancing the state of the art in this domain.

3.3. Summary

This chapter reviews the state-of-the-art models for FER, SER, and Textual Emotion Analysis, which were explored to develop a functional and testable prototype for the INTU-AI system. Publicly available pre-trained models were leveraged for each modality.

For FER, the system incorporates Face Detection RFB-320 and a VGG13 model trained on the FER+ dataset. FER+ enhances facial expression classification by introducing refined annotations and additional emotion classes. The VGG-Face model, originally designed for large-scale face recognition, and the WIDER FACE dataset were also instrumental in enabling robust face detection under diverse real-world conditions.

The SER module relies on a fine-tuned version of the XLSR-53 wav2vec2 model trained on the RAVDESS dataset. RAVDESS provides high-quality emotional audio-visual data, enabling accurate speech-based emotion detection.

For Textual Emotion Analysis, the AffectAlchemy dataset, based on Plutchik's Wheel of Emotions, was employed to train various classical models (e.g., SVM, Logistic Regression), and deep learning, CNN and RNN with GloVe/Fasttext embeddings, models. Additionally, state-of-the-art transformer models such as RoBERTa and DistilBERT were fine-tuned for improved emotion classification.

For the second part of this work, we embrace in the multimodal domain, the MELD dataset was chosen for experimentation due to its integration of audio, visual, and textual data in emotionally annotated dialogues. As a benchmark, the GraphSmile model was reviewed for its innovative use of graph structure fusion and sentiment dynamics perception in multimodal conversations.

In summary, this chapter outlines the theoretical foundation and empirical resources that support the multimodal and deception-aware architecture proposed in this work, emphasizing the relevance and innovation of integrating emotion recognition into investigative frameworks.

4. Methodology

The dual nature of this work, one part focused on the development of an application and the other on the development of a concept, necessitated the use of distinct methodologies tailored to each facet of the problem.

Since both components aim to address a client-centred issue, namely providing a framework to assist PJM investigators during the interrogation process, the selected methodologies are closely aligned with project execution best practices. Specifically, the first part of the work follows a combination of evolutionary prototyping and CRISP-DM [70], while the second part adopts CRISP-DM exclusively as its guiding methodology.

4.1. Cross-Industry Standard Process for Data Mining

The methodology adopted for this work is based on the CRISP-DM approach (Cross-Industry Standard Process for Data Mining), an industry-independent process model widely used in data science projects. CRISP-DM is structured into six iterative phases, ranging from business understanding to deployment. The image below illustrates the cyclical process underlying the methodology adopted in this study.

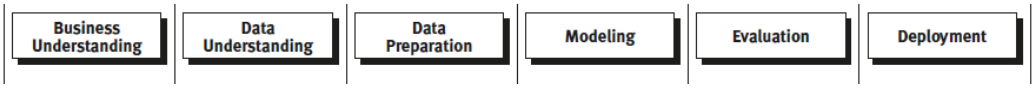


Figure 7 - CRISP-DM reference model life cycle (source [70])

Given that this project is divided into two distinct yet interconnected areas, it became necessary to adapt the methodology accordingly. While the Business Understanding phase is shared across both parts of the project, the remaining phases differ due to the use of different datasets and data science techniques. Therefore, the methodology has been tailored to reflect this structure. The Business Understanding phase is detailed in Annex D – Business Understanding.

Data understanding, preparation, and modelling were addressed both during the initial development of the INTU-AI program, particularly in the construction of the emotion-from-text model, as well as in the conceptual design of the multimodal models. The subsequent stages, are addressed separately in the end of the work, since each chapter deals with different datasets and applies different techniques and models within the domain of data science. Evaluation and deployment activities are integrated throughout the test and validation cycle.

4.2.Evolutionary Prototyping

To transform PJM/IA program from an initial concept to the program that supports PJM investigators during interrogations, we adopted an Evolutionary Prototyping model, which enabled incremental system development through iterative cycles of continuous improvement and ongoing validation by end-users.

The process began with a detailed requirements phase, which included conducting interviews with PJM investigators to understand operational needs, a summary of which is available in Annex B – User Story: INTU-AI. This phase also involved defining the Operational Requirements and Technical Specifications, documented in Annex C – Operational Requirements/Technical Specifications for INTU-AI. These activities were carried out through in-person meetings and a thorough review of existing documentation.

During the Evolutionary Prototyping cycle, which included phases of design, prototyping, customer evaluation, and review and refinement, biweekly meetings were held to align details between the PJM and the development team. One of the key assumptions established during this phase was that the program would serve as a prototype, developed using Python for both backend and frontend, primarily due to the team's familiarity with the technology. However, in the near future, the system may be migrated to frameworks that offer improved user-friendliness and broader usability.

The evolutionary prototyping model allowed for the creation of incremental versions of the system, which were continuously validated by the client throughout the project. Proof-of-concept demonstrations were carried out using both synthetic and real interrogation data. Each iteration of the prototype was refined based on direct user feedback, following a human-in-the-loop validation approach, which ensured ongoing alignment with operational needs and usability expectations.

The testing and release phases took place in February 2025, during which version 1.0 was delivered to the PJM. The program is currently being tested and applied in both real and simulated cases within the institution.

5. INTU-AI application

The INTU-AI program was developed with both its backend and frontend components implemented in Python. The main objective of the system is to support the digitalization of the interrogation process within the PJM. As output, the tool provides the user with emotion-labelled video or audio files, generated through emotion recognition models, and additionally supports the automatic generation of the necessary interrogation reports. As an input, the interrogator is only required to provide an identifier for the individual being interrogated and either an audio/video file or to conduct a live interrogation session.

5.1. Program Architecture

The INTU-AI program is a single-user application operated through a central interface. It is based on the provision of a data set to the system, which then processes the input and returns a collection of reports and annotated videos to support the interrogator's analysis. The architecture of the system is illustrated in the following figure.

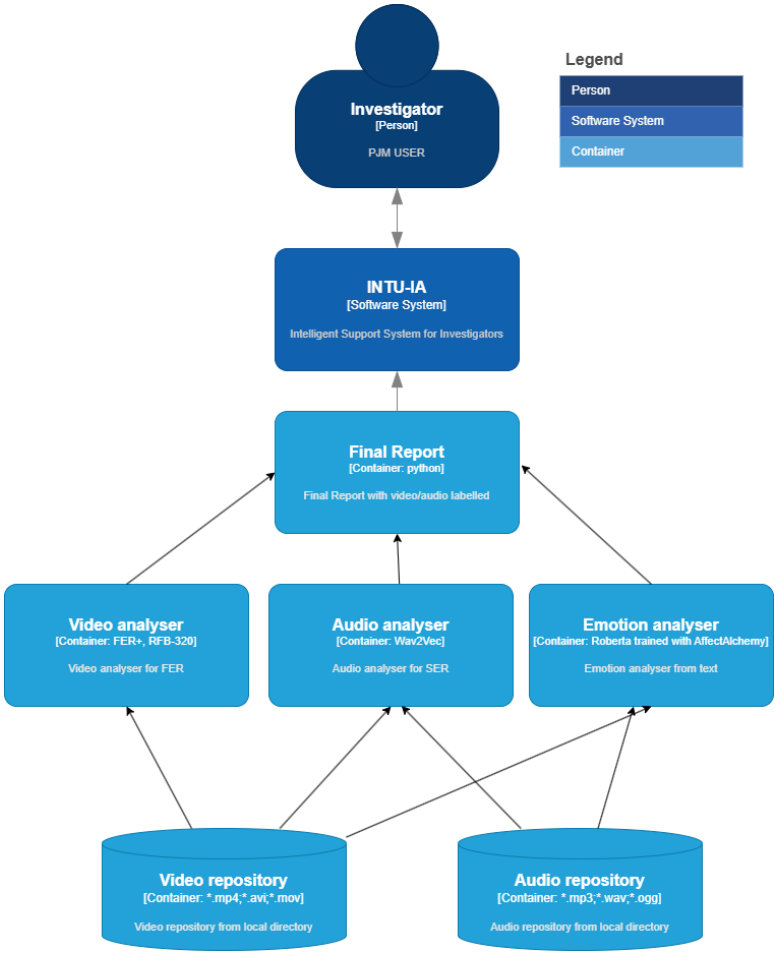


Figure 8 - C4 Model for INTU-AI

5.2. Graphical Interface

The graphical interface is built using the Tkinter and CustomTkinter libraries. It begins with an authentication system which, upon successful validation, grants access to the main window. This window is divided into four sections, three of which are informative and one interactive. The first quadrant, located in the top-right corner, contains a set of combo boxes, text boxes, and buttons for uploading documents and accessing features such as viewing reports. The second and third quadrants, positioned in the bottom-right and bottom-left corners respectively, display graphical information related to emotional analysis during the interrogation, where the results of SER and FER are presented. The fourth quadrant, located in the top-left corner, presents personal information about the interrogated individual.

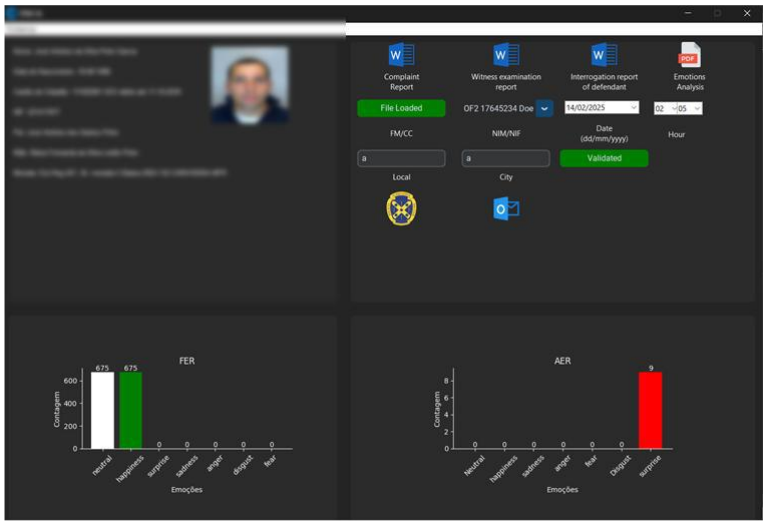


Figure 9 - Completed main menu

At the outset, the user provides the system with two input files: a PDF document containing the identification details of the individual being interrogated, and a corresponding video or audio file. This media file may either be uploaded manually by the interrogator or captured in real time through direct recording, enabling the system to be used both in retrospective analysis and during live interrogations.

The following diagram aims to summarise the entire process related to the user's interaction with the INTU-AI system, detailing how the inputs provided by the user are processed internally by the application.

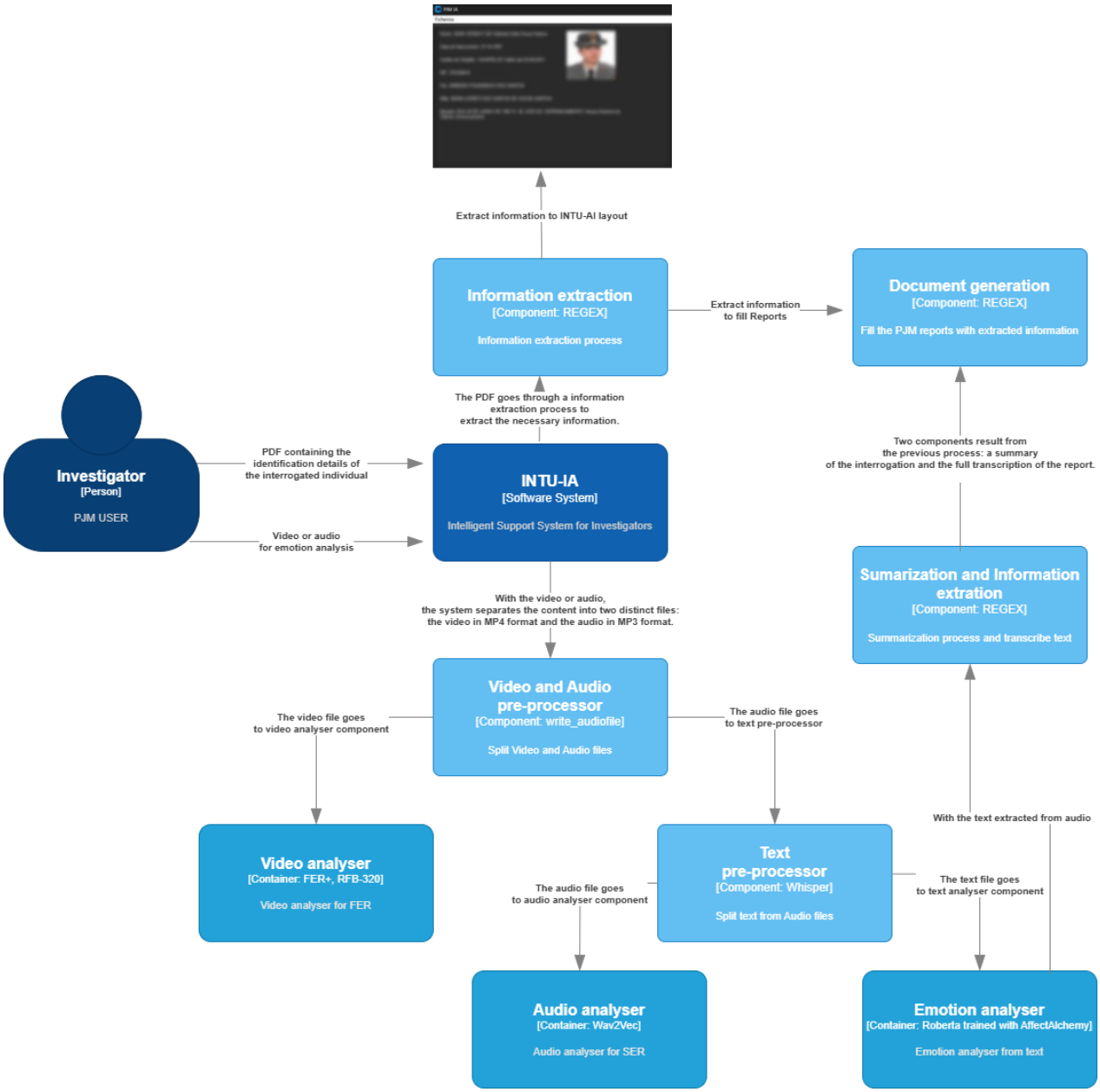


Figure 10 - C4 model for graphical interface

During the initial processing of the PDF, the system applies regular expressions and custom-designed pattern-matching techniques tailored to the specific context, combined with information extraction methods, to retrieve the necessary data. This extracted information is then used to populate two key areas: the previously mentioned interface quadrant that displays relevant personal details of the interrogated individual, and the pre-filling of official reports currently in use by the PJM.

As PDF input, the user may choose to upload either a scanned copy of the national identification card or a military registration form, the latter being the most commonly used,

as it serves as the standard identification document for military personnel. This process is illustrated in the following figure.

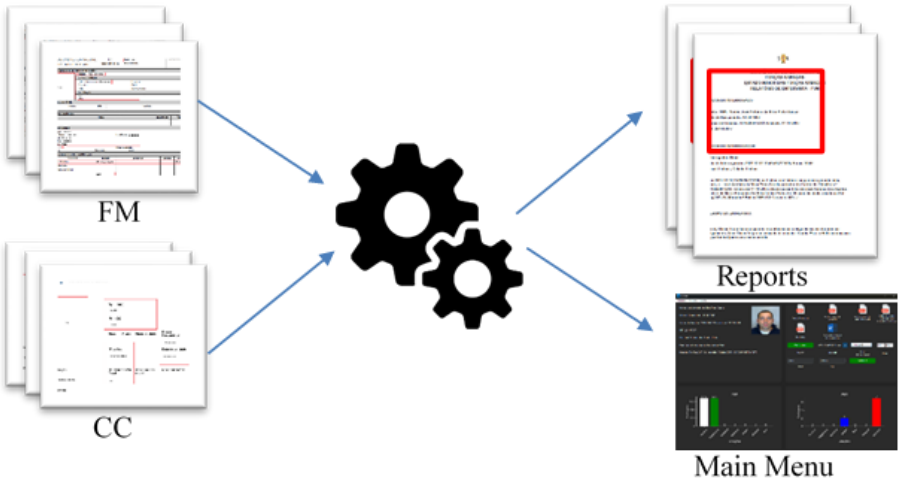


Figure 11 - Flowchart of the information extraction process in the project

When an audio or video file is introduced, the initial step of the system consists of separating the file into its analysable components, in accordance with the three emotion analysis vectors: video for FER, audio for SER, and text for text-based emotion analysis. The user has three available input options: providing a video file, which inherently includes video, audio, and derived text; providing an audio-only file, which includes both audio and extracted text; or using a live video stream captured via camera, which, in practice, is treated by the system as a standard video input.

This process feeds into the subsequent stage of sentiment analysis across the three vectors, in cases where the user has provided a video file or used the open camera option. Alternatively, if the user has chosen to analyse only an audio file, such as a recorded surveillance audio, the analysis will be limited to two vectors: audio and text.

During the text extraction process from the audio component, the use of Whisper Large enables a level of detail sufficient to serve as a verbatim transcription, which can be directly integrated into the official PJM report. This transcription is further complemented by a summarisation, also intended to support the automatic completion of the report.

Still within the graphical interface, the user has access to a set of optional fields displayed in the second quadrant of the main panel. These fields are complementary to the automatic report generation but are not mandatory. They are intended for situations in which the

interrogator wishes to manually enter information different from the default values set during the initial phase. For example, the location and city where the interrogation took place may be adjusted, since interrogations are often conducted not at the PJM headquarters in Lisbon, but rather at various military units across the country.

In the following figure, the previously mentioned options can be observed. Additionally, at the top section of the menu, there are three Word documents and one PDF file displayed. The Word documents correspond to the official templates currently in use by the PJM and indicate the specific type of interrogation to be conducted. By default, INTU-AI automatically fills in all three templates and makes them available to the user. Lastly, the PDF file, titled "Emotions Analysis", serves as the final report, compiling a set of relevant information intended to support the interrogator as further detailed in a later section.

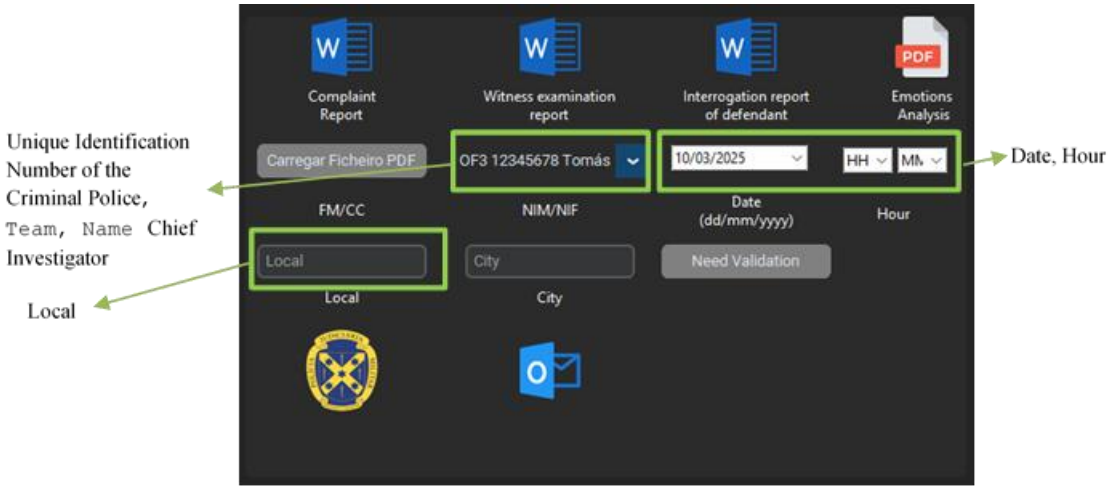


Figure 12 - First quadrant, interactive menu

5.3.Emotion analysers components

The components used align with the requirements for emotion analysis, specifically covering FER, SER, and text-based emotion analysis. For the FER module, we employed pre-trained models, namely the Face Detection RFB-320, trained on the WIDER FACE dataset for face detection, and a VGG13 model trained on the FER+ dataset for facial emotion classification, adapted from [59]. For the audio component, we used a fine-tuned version [63] of the XLSR-53 large model [64], which was also pre-trained.

As for the text component, several models were trained using the AffectAlchemy dataset, and the model with the best classification performance was selected; this aspect is discussed in detail in a section ahead.

According to Annex C – Operational Requirements/Technical Specifications for INTU-AI, emotions must be analysed in 10-second intervals, and the emotion to be returned should be the most frequently observed within each time window. To ensure compliance with this requirement, both video and audio inputs are divided during the Video and Audio Pre-processing stage into 10-second segments. Each of these clips is then labelled with the corresponding dominant emotion.

To preserve the logical sequence of the original content, the clips are named in a way that reflects their temporal position, which is also maintained throughout the subsequent processing components. At the end of the emotion analysis process, the user receives two outputs: a final video annotated with the recognised emotions from FER, SER, and text-based analysis, and an annex to the final report containing a time-aligned table indicating the dominant emotion detected in each 10-second segment. The final video is produced by concatenating all labelled 10-second clips, thus preserving the chronological flow while visually presenting the emotional data.

As output, the INTU-AI system provides the interrogator with a folder stored locally on their device. This folder contains the documents related to the identification of the interrogated individual, the complete video annotated with the predominant emotion detected in each 10-second interval, and the official PJM report templates duly filled in, as previously described. Additionally, the folder includes a final report composed of a main body and four annexes.

The main body contains the identification of the interrogator, as well as the identification of the interrogated individual, which is extracted from the submitted documents. Annex A includes the full transcription of the interrogation or surveillance audio, accompanied by a summary of its contents. Annexes B, C, and D present the written results of the emotion analysis corresponding to each of the three vectors: FER, SER, and text-based emotion analysis. The following figure summarises the entire process described above.

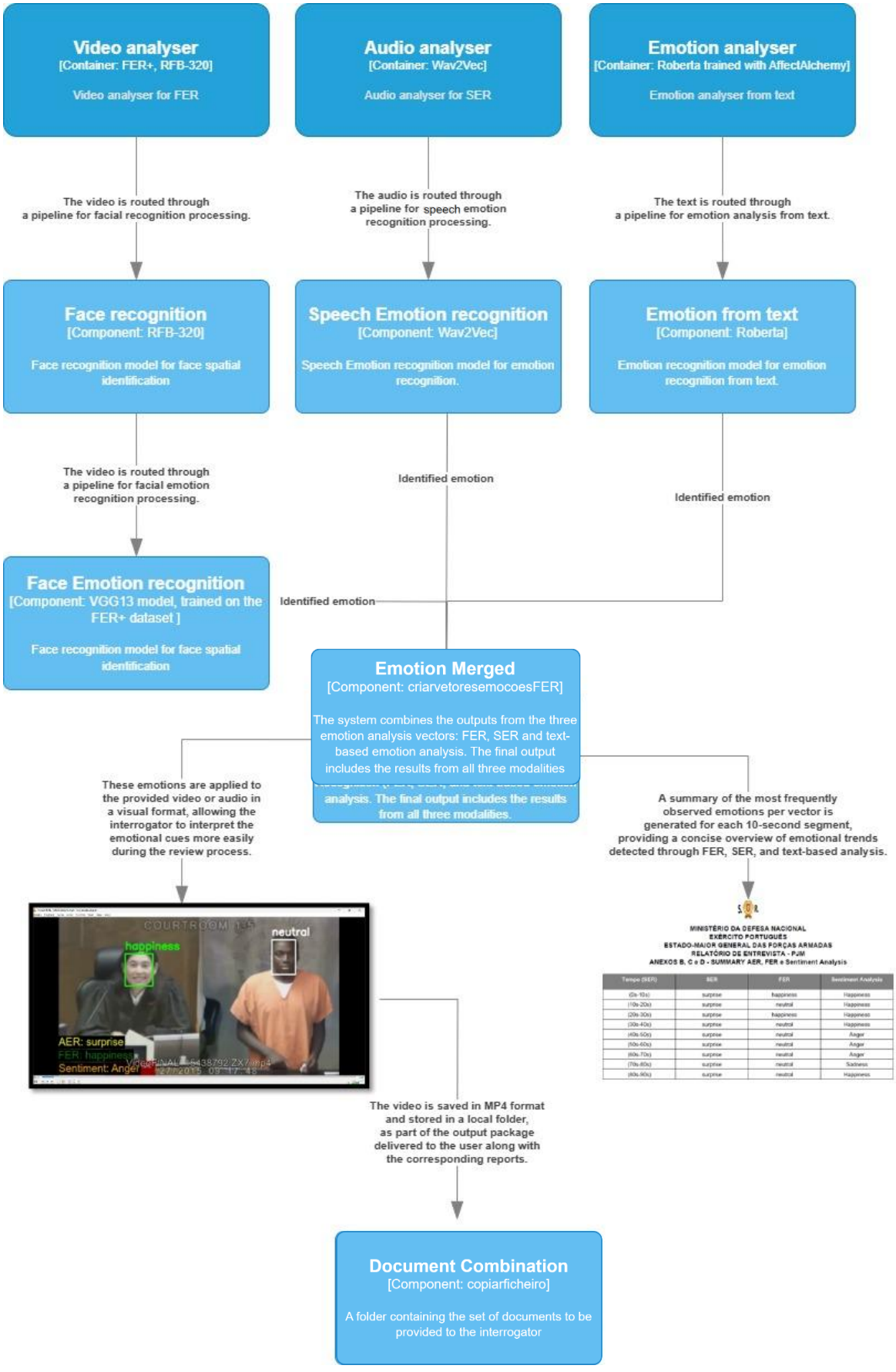

Video analyser
[Container: FER+, RFB-320]
Video analyser for FER
Audio analyser
[Container: Wav2Vec]
Audio analyser for SER
Emotion analyser
[Container: Roberta trained with AffectAlchemy]
Emotion analyser from text
The video is routed through a pipeline for facial recognition processing.
The audio is routed through a pipeline for speech emotion recognition processing.
The text is routed through a pipeline for emotion analysis from text.
Face recognition
[Component: RFB-320]
Face recognition model for face spatial identification
Speech Emotion recognition
[Component: Wav2Vec]
Speech Emotion recognition model for emotion recognition.
Emotion from text
[Component: Roberta]
Emotion recognition model for emotion recognition from text.
The video is routed through a pipeline for facial emotion recognition processing.
Identified emotion
Identified emotion
Face Emotion recognition
[Component: VGG13 model, trained on the FER+ dataset]
Face recognition model for face spatial identification
Identified emotion
Emotion Merged
[Component: criarvetoresemocoesFER]
The system combines the outputs from the three emotion analysis vectors: FER, SER and text-based emotion analysis. The final output includes the results from all three modalities
These emotions are applied to the provided video or audio in a visual format, allowing the interrogator to interpret the emotional cues more easily during the review process.
A summary of the most frequently observed emotions per vector is generated for each 10-second segment, providing a concise overview of emotional trends detected through FER, SER, and text-based analysis.
COURTROOM 1
happiness
neutral
AER: surprise
MINISTÉRIO DA DEFESA NACIONAL
EXÉRCITO PORTUGUÊS
ESTADO-MAIOR GENERAL DAS FORÇAS ARMADAS
RELATÓRIO DE ENTREVISTA - PJM
ANEXOS B, C e D - SUMMARY AER, FER e Sentiment Analysis
The video is saved in MP4 format and stored in a local folder, as part of the output package delivered to the user along with the corresponding reports.
Document Combination
[Component: copiarficheiro]
A folder containing the set of documents to be provided to the interrogator

Figure 13 - C4 model for Emotion analysers components

5.3.1. Emotion Analysis from text

As discussed in the chapter dedicated to the literature review, we used the AffectAlchemy dataset [65] to train our Emotion Analysis model. This dataset is available in CSV format on the GitHub repository provided by the authors [67].

During the development of this work, the initial idea was to build a model capable of receiving input in Portuguese and classifying it accordingly. However, we encountered a significant challenge: the lack of well-established benchmark datasets for the Portuguese language in the context of emotion and sentiment analysis.

To address this limitation, we adopted the following approach: the model development was divided into two separate pipelines, one using a dataset in English ("*Modelo Sentiment analises.ipynb*") and the other using a dataset in Portuguese ("*Modelo Sentiment analises_PT.ipynb*"). This allowed us to not only evaluate performance using established English-language benchmarks, but also to compare the results between the two models (PT *vs.* EN), providing insights into the impact of language on model performance and robustness.

5.3.2. Understanding the Data

The dataset contains 20,075[15] observations and two features ("Text" and "Emotion"), with the text variable consisting of several sentences, mostly in English. The first step was to ensure that all sentences were written in English. Then, we analysed the distribution of emotions throughout the dataset as represented at the next table.

Table 3 - Emotion count dataset AffectAlchemy [67].

Emotion	Count	Emotion	Count	Emotion	Count
Joy	1723	Gratitude	1030	Grief	233
Sad	1577	Disgust	1028	sadness	3
Fear	1490	Contempt	1011	love	3
Happy	1324	Love	1005	Shame, Sad	2
Surprise	1297	Anticipation	969	Gratitude, Love	2
Neutral	1214	Awe	969	fear	2
Peace	1132	Optimism	759	surprise	2
Determination	1058	Trust	625	Peace, Joy	1
Anger	1043	Shame	572	Shame, Fear	1

[15] After excel pre-processing (initial 20,083)

In the emotion analysis, we are only interested in the emotions addressed by the other two models, namely: “Sadness”, “Happiness”, “Disgust”, “Anger”, “Fear”, “Surprise”, and “Neutral”. Additionally, since this dataset was built on the premise that a sentence may not correspond to a single emotion but rather to a pair of emotions, as proposed by Plutchik [66], this aspect is also taken into consideration.

After applying the indicated modifications, the dataset was reduced to a total of 8,988 observations, the following images represent the dataset after the previously applied modifications.

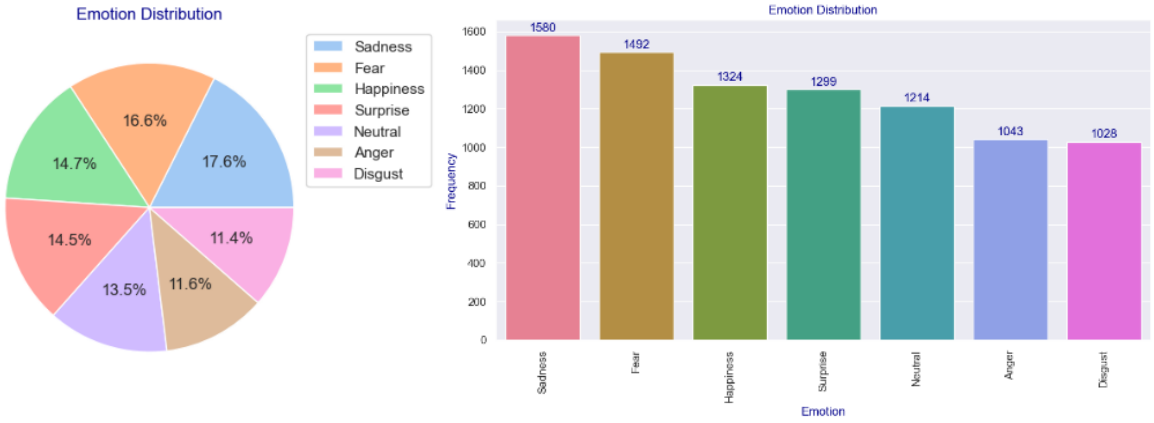


Figure 14 - Pie chart and Bar chart of emotion distribution

From the graphical and tabular information presented above, we can conclude that our dataset is imbalanced, with the emotion sadness having the highest number of observations and disgust the lowest.

Regarding the named entities present in the dataset, a total of eighteen distinct entity types were identified. Their absolute and relative distributions are presented in the table below.

Table 4 - Table of NER from the dataset

NER	Count	%	NER	count	%
CARDINAL	1592	25.50	MONEY	81	1.30
PERSON	1229	19.69	PRODUCT	43	0.69
DATE	1210	19.38	FAC	37	0.59
ORG	736	11.79	LOC	37	0.59
TIME	403	6.46	QUANTITY	35	0.56
GPE	366	5.86	EVENT	28	0.45
ORDINAL	168	2.69	PERCENT	21	0.34
NORP	147	2.35	LANGUAGE	15	0.24
WORK_OF_ART	89	1.43	LAW	6	0.10

The NER analysis of the dataset revealed a significant number of mentions across various entity categories, with some appearing more frequently than others. The first four categories alone account for approximately 76% of all recognized entities, highlighting a strong presence of quantitative references, people, dates, and organizations within the textual content. Other notable categories include TIME (6.43%), GPE (geopolitical entities 5.95%), and ORDINAL (ordered numbers 2.69%). Less frequent categories, such as LAW, EVENT, PERCENT, and LANGUAGE, each represent less than 1% of the total entity mentions, indicating a residual presence in the dataset. Overall, the data indicates that the textual corpus is predominantly composed of numerical information and references to people, dates, and organizations, which may be relevant for more in-depth semantic analysis or for enriching linguistically-driven machine learning models.

To gain a general understanding of the emotional tone expressed in the dataset, a sentiment analysis was performed using the TextBlob library. This analysis returned the following overall values, polarity with 0.092[16] and subjectivity with 0.536[17]. The polarity score ranges from -1 (completely negative) to +1 (completely positive). A value of 0.092 indicates a slightly positive sentiment, suggesting that, on average, the texts in the dataset lean marginally toward a positive tone. The subjectivity score ranges from 0 (completely objective) to 1 (completely subjective), a value of 0.536 reflects a moderate degree of subjectivity, meaning that the texts tend to contain personal opinions, emotions, or subjective expressions, rather than purely factual or objective information. In summary, the dataset shows a slightly positive and moderately subjective emotional profile, which aligns with the nature of user-generated content, where emotions and opinions are frequently expressed.

5.3.3. Data Preparation

In order to ensure the quality and consistency of the textual data used in this work, a comprehensive text pre-processing pipeline was applied to the dataset. This process aimed to clean, normalize, and prepare the text for subsequent natural language processing and machine learning tasks. The steps applied are described below[18]:

[16] Polarity - 0.106 with the dataset in Portuguese.
[17] Subjectivity - 0.696 with the dataset in Portuguese.
[18] The same type of processing was applied to the Portuguese dataset, with the necessary adaptations made to account for the specific characteristics of this dataset.

- Using LanguageTool[19], we corrected various grammatical and contextual errors in our dataset.
- Regarding text cleaning and normalization, each text entry was subjected to several cleaning operations which integrated NLP tools such as spaCy[20], NLTK[21], TextBlob[22], and NeatText[23].
- Tokenization and lemmatization, after cleaning, the text was lemmatized using spaCy's large English model[24], extracting only alphabetical tokens and removing stop words. Lemmatization ensures that different word forms are reduced to their root form, which enhances consistency in downstream analysis.
- Hashtag decomposition, custom handling of hashtags was implemented, which splits compound hashtags.
- NER and anonymization, named entities such as persons, dates, organizations, and cardinals were extracted via spaCy and optionally removed/anonymized to preserve privacy and reduce potential bias in the model learning process. This anonymization was conducted by deleting identified spans from the raw text.
- POS tagging, each token in the cleaned text was annotated with its POS tag using spaCy, enabling deeper syntactic understanding of sentence structure and serving as an additional feature for downstream classification models.
- Sentiment scoring and Sentiment polarity were calculated for each cleaned text using VADER (Valence Aware Dictionary and Emotion Reasoner) [71], assigning a compound score between -1 (negative) and +1 (positive) to each instance.
- Token extraction, the final cleaned text was tokenized into individual words using NLTK's word tokenize, producing a list of tokens per instance.

As a result of the pre-processing steps described above, the dataset was reduced to 8016 observations, as some entries became empty after cleaning. This also led to a reduction in the total number of words to 57619, representing approximately 11% decrease in the overall word count. Following the text pre-processing phase, a sentiment analysis was conducted on

[19] Available at https://languagetool.org/
[20] Available at https://spacy.io/
[21] Available at https://www.nltk.org/
[22] Available at https://textblob.readthedocs.io/
[23] Available at https://jcharis.github.io/neattext/
[24] Available at https://spacy.io/models/en#en_core_web_lg**,** for Portuguese was used the small model https://spacy.io/models/pt#pt_core_news_sm

the cleaned dataset using the TextBlob library. The overall sentiment values obtained were the polarity scores 0.056 and subjectivity 0.559[25].

The polarity score, indicates a slightly positive sentiment across the dataset. The subjectivity score, suggests a moderate level of subjectivity, reflecting the presence of personal opinions and emotional expressions in the text. These results confirm that, even after data cleaning, the dataset maintains a mildly positive and moderately subjective tone, which is consistent with the nature of emotionally expressive textual content.

5.3.4. Modelling

Models were trained using SVC, Naïve Bayes, Random Forest, Decision Tree, XGBoost, KNN, and Logistic Regression. We also explored two deep learning approaches: a CNN with pre-trained GloVe embeddings, and a RNN also using GloVe embeddings.

Moreover, we further enhanced our analysis by fine-tuning state-of-the-art pre-trained transformer models, namely DistilBERT [52] and RoBERTa [36], using our dataset. The performance metrics of the proposed models are summarized in the table below.

Table 5 - Performance metrics summary of the models on the dataset validation set.

	English				Portuguese			
Model/Metrics	**Acc**	**F1**	**Recall**	**Prec**	**Acc**	**F1**	**Recall**	**Prec**
Regression Log.	0.6721	0.6720	0.6721	0.6737	0.6481	0.6482	0.6481	0.6490
NB	0.6652	0.6714	0.6652	0.6947	0.6418	0.6468	0.6418	0.6646
XGBoost	0.6428	0.6435	0.6428	0.6465	0.6085	0.6128	0.6085	0.6190
DT	0.4882	0.4826	0.4882	0.4809	0.4774	0.4756	0.4774	0.4755
RF	0.6297	0.6283	0.6297	0.6451	0.5853	0.5871	0.5853	0.6035
KNN	0.3828	0.3899	0.3828	0.5888	0.4849	0.4737	0.4849	0.4853
SVC	0.6690	0.6713	0.6690	0.6778	0.6255	0.6288	0.6255	0.6359
CNN + GLoVe	0.5842	0.5836	0.5842	0.5900	0.5602	0.5602	0.5602	0.5667
RNN + GLoVe	0.6172	0.6091	0.6172	0.6130	0.5941	0.5872	0.5941	0.5972
Roberta	**0.7519**	**0.7526**	**0.7519**	**0.7558**	**0.6888**	**0.6916**	**0.6888**	**0.7043**
DistilBERT	0.7182	0.7201	0.7182	0.7320	0.6637	0.6671	0.6637	0.6861

In comparative terms, the transformer-based model RoBERTa demonstrated clear superiority in both languages, English and Portuguese. Given the imbalanced nature of our dataset, the F1-score was selected as the primary evaluation metric. The RoBERTa model

[25] Polarity - 0.107 and Subjectivity - 0.493 with the dataset in Portuguese.

fine-tuned during our experiments is the one used to classify emotions resulting from the text-based emotion analysis component within the INTU-AI system.

5.3.5. Testing Emotion Analysis from text – AffectAlchemy dataset

At the test set level, the models evaluated achieved results that were very similar to those obtained on the validation set, with the RoBERTa model consistently outperforming the others across all evaluation metrics. The table below presents the performance results on the AffectAlchemy dataset for the test set.

Table 6 - Performance metrics summary of the models on the dataset test set.

	English				Portuguese			
Model/Metrics	**Acc**	**F1**	**Recall**	**Prec**	**Acc**	**F1**	**Recall**	**Prec**
Regression Log.	0.6640	0.6637	0.6640	0.6679	0.6631	0.6648	0.6631	0.6701
NB	0.6359	0.6405	0.6359	0.6637	0.6499	0.6530	0.6499	0.6692
XGBoost	0.6222	0.6260	0.6222	0.6341	0.6167	0.6196	0.6167	0.6274
DT	0.4944	0.4944	0.4944	0.4957	0.5094	0.5075	0.5094	0.5099
RF	0.6284	0.6285	0.6284	0.6495	0.5891	0.5878	0.5891	0.6061
KNN	0.3747	0.3783	0.3747	0.5863	0.5151	0.5094	0.5151	0.5247
SVC	0.6515	0.6532	0.6515	0.6613	0.6330	0.6361	0.6330	0.6449
CNN + GLoVe	0.5723	0.5700	0.5723	0.5768	0.5690	0.5687	0.5690	0.5734
RNN + GLoVe	0.5829	0.5759	0.5829	0.5816	0.5928	0.5876	0.5928	0.5951
Roberta	**0.7182**	**0.7142**	**0.7182**	**0.7139**	**0.6738**	**0.6715**	**0.6738**	**0.6722**
DistilBERT	0.6833	0.6842	0.6833	0.6928	0.6274	0.6270	0.6274	0.6474

In parallel, and considering the replication of the authors' work [65], carried out during the preliminary phase of this project where we had achieved an accuracy of 0.6618 and an F1-score of 0.6655, it is important to note that the dataset available on GitHub contains approximately 10 000 fewer observations than what is stated in the original study, representing a reduction of about one-third of the data. Our proposed model was applied following the same procedure described by the original authors, and the results are presented in the table below.

Table 7 - Comparison Between Our Work and the Authors' Work

Model/Metrics	**Acc**	**F1**
Initial metrics from the replication of the authors' work	66.18	66.55
Ours study	**85.82**	**85.72**
Difference	**19.64**	**19.17**

We can assert that our model increased the dataset's classification capacity by 19.17% compared to replicating the authors' work. More sophisticated transformer-based models, as RoBERTa, significantly outperform traditional methods in English, but this advantage

diminishes in the Portuguese context. The performance gap between English and Portuguese suggests that many English pre-trained resources do not adequately capture Portuguese's semantic particularities, thereby requiring additional adaptations, more extensive fine-tuning, or even dedicated data collection for this language.

5.4. Output flow

As the final output, the user has access to five documents, in addition to the previously mentioned Word files, which represent the official report templates required by this police force, pre-filled with the information extracted throughout the process. The interrogator is also provided with a compiled video, annotated with the most frequently observed emotions in each 10-second interval. The following figure aims to summarise the entire process carried out by the program, complementing the points discussed above.

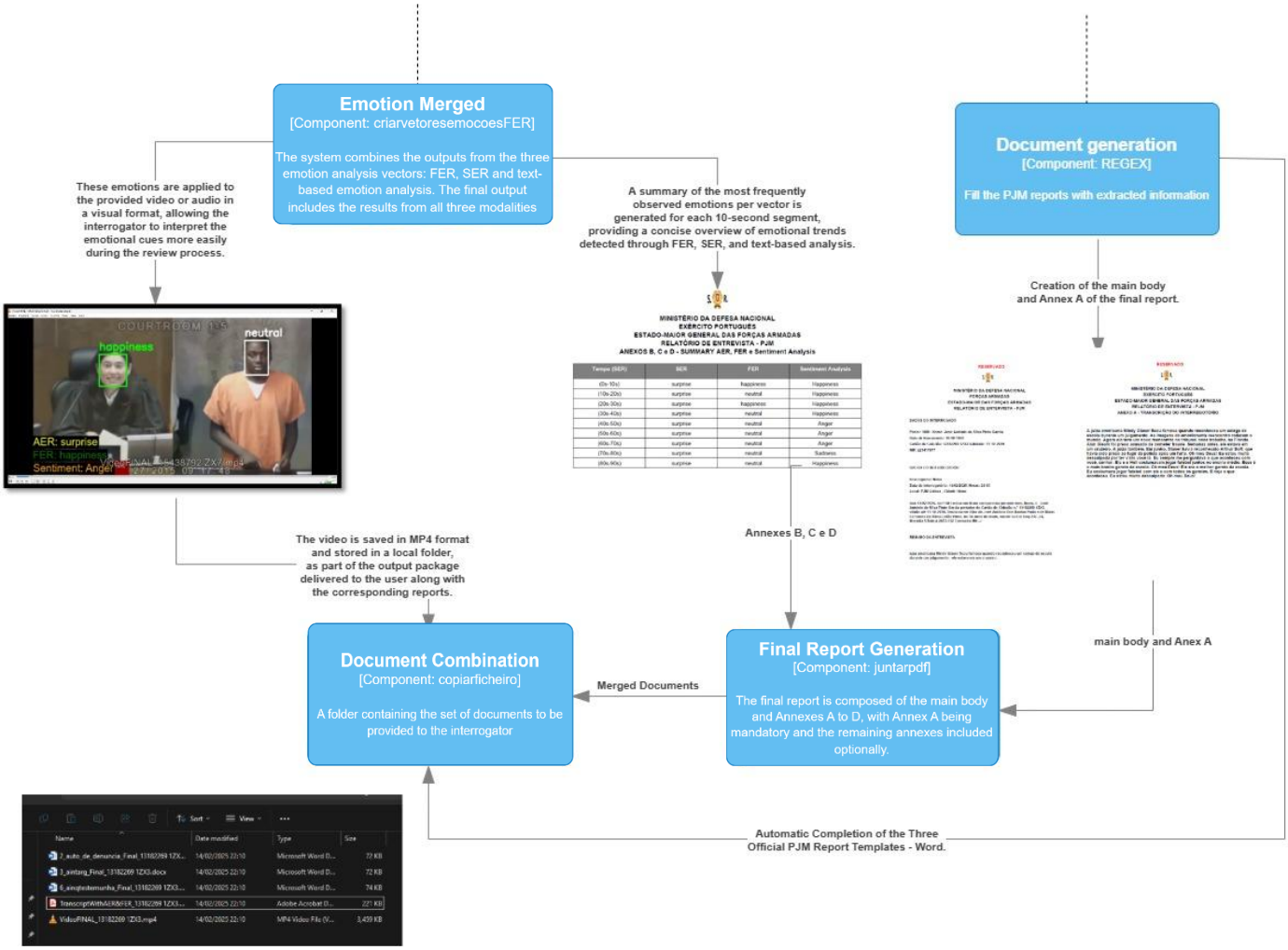


Figure 15 - C4 model for output flow

Additionally, as a summary of the entire process, the interrogator receives a final report which, beyond including the identification details of both the interrogator and the

interrogated individual in the main body, features an annex containing the full transcription of the input provided to INTU-AI, along with a set of annexes summarising the emotions identified across the analysed vectors over time.

5.5. Testing INTU-AI

One of the main challenges faced during the development of the program was the lack of available interrogation videos to properly test the system and make the necessary refinements that every program requires.

Due to the confidential nature of real interrogations and compliance with General Data Protection Regulation (GDPR) regulations, using actual videos for testing was not an option. To overcome this limitation, we adopted a different strategy: creating fictional interrogation videos using open-source AI tools, specifically ChatGPT [21] and InVideo AI [79]. In practice, this platform converts text into video, allowing us to simulate realistic interrogation scenarios for testing purposes.

Figure 16 - Video Creation Workflow for Testing

Final testing of the system took place during the PJM inspector training course in February and March 2025. Additional testing remains ongoing, particularly in the areas of performance, compatibility, security, and data confidentiality. In terms of final documentation and training, a bilingual user manual (Portuguese and English) was delivered Annex A- INTU-IA - User Guide - VERSION 1.0. An initial training session was provided to the PJM unit responsible for testing, with additional sessions scheduled for the remaining personnel.

The software was officially delivered at the end of January 2025 in an executable .exe format. The project is currently in its initial support phase, with ongoing corrective maintenance and planned improvements based on feedback.

5.6. Summary

This chapter presents the design and implementation of INTU-AI, a Python-based application developed to support the PJM by digitising and automating administrative procedures related to interrogations. The system integrates three core emotion analysis components FER, SER, and text-based emotion analysis forming a complete end-to-end solution that processes inputs and generates structured reports.

INTU-AI is a single-user system with a graphical interface built using Tkinter and CustomTkinter, it accepts as input a PDF document with identification details and a video or audio file, or alternatively a live video stream. The PDF is processed using regular expressions and information extraction methods, while the media file undergoes pre-processing to segment the content into 10-second intervals for analysis.

For FER, the system uses pre-trained models such as Face Detection RFB-320 and VGG13. The SER component relies on a fine-tuned XLSR-53 model, and for text analysis, a custom-trained RoBERTa model, built using a curated version of the AffectAlchemy dataset, was selected due to its superior F1-score performance. The RoBERTa model stood out for its classification performance when compared to other models tested during the modelling phase. These included traditional machine learning algorithms such as SVC, Naïve Bayes, Random Forest, Decision Tree, XGBoost, KNN, and Logistic Regression, as well as deep learning approaches like CNN and RNN using GloVe and FastText embeddings.

As final output, the user receives a local folder containing: the three pre-filled official PJM report templates in Word format; a video annotated with the most frequent emotions in 10-second intervals; and a PDF report with identification data, a full transcription and summary (Annex A), and written results for each emotion analysis vector (Annexes B, C, and D). This solution aims to enhance the interrogation process by combining automation, structured reporting, and multimodal emotional insight.

6. Multimodal Integration

Multimodal integration arises primarily from the need to provide the interrogator with a unified emotional assessment that is, in cases where the detected emotions differ across modalities, to determine which emotion is truly predominant. According to numerous emotion researchers, facial expression is considered the most immediate and dominant channel of emotional communication among humans, a view extensively studied and validated in foundational works such as those by Ekman [72, 73], Izard [74], Keltner [75], Darwin [76], among others.

This perspective is particularly relevant in light of advances in machine learning, where it becomes possible to conduct a proof of concept to investigate whether facial emotion indeed prevails over other modalities in an interrogation context. Another line of inquiry is whether all modalities contribute equally to the emotional inference, or conversely, if certain modalities carry greater relative importance. To address this, we explored the available options applicable to this case:

1. Facial emotion predominates over other modalities in emotion analysis.
2. All modalities have equal relative importance and should be analysed in a unified manner.
3. The modalities, vectors, contribute with different relative importance to the overall emotional interpretation.

The lack of publicly available police interrogation datasets limits our ability to apply this proof of concept in a way that fully mirrors real-world scenarios. Consequently, the choice of the MELD dataset represents the most suitable alternative, as it closely approximates real-life contexts. Unlike simulated environments, MELD captures natural conversational settings, including interactions that, although embedded in a comedic context, often involve elements of deception, manipulation, or dishonesty.

6.1. Understanding the Data

After obtaining the dataset [77], all video files were consolidated into a single unified folder, resulting in a total of 13708 video files, distributed as follows in Figure 17.

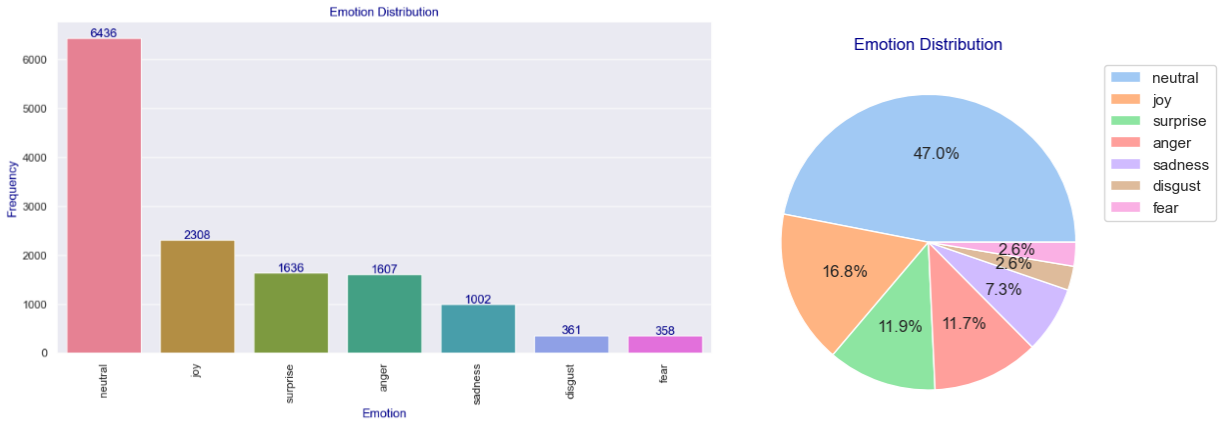


Figure 17 - Bar plot and Pie plot chart MELD without cleaning

Given the multimodal context involving three distinct vectors, it is essential to thoroughly understand the data associated with each modality, video, audio, and text. This understanding is crucial for the effective design and development of the model. Therefore, the following analysis is divided into three parts: video, audio, and text.

To achieve this, we employed a pipeline to extract both the audio and the corresponding transcriptions from the original video files. For audio extraction, we used a native Python utility to convert .mp4 files into .wav format. Subsequently, we utilized a pipeline based on the Whisper large model [79] to transcribe the audio into text.

6.1.1. Video data understanding

In the analysis of the 13708 video samples, we observed that the average duration was approximately 3 seconds. However, upon examining the non-central tendency measures, we identified an anomalous case, the video 1165_neutral, which had a recorded duration of 0 seconds. Such a duration was deemed invalid for any of the intended modalities (video, audio, or text), and therefore, this sample was excluded from the dataset to ensure data integrity and consistency.

The average number of frames in our videos is approximately 75, with the frame rate predominantly set at 23.98 fps. Only 67 instances exhibit a frame rate of 25 fps; however, this discrepancy is not considered significant and may be attributed to the videos having

been previously trimmed. In terms of resolution, 99.5% of the videos have a resolution of 1280x720. The remaining 0.5% are mostly associated with a resolution of 496x384 (69 instances), and a smaller portion with 560x432 (7 instances).

If all videos with divergent frame counts and resolutions were associated exclusively with the majority class, their elimination could be considered justified. However, since these videos contain information pertaining to minority emotion classes, their presence is valuable and, therefore, they were retained. As we can see in the chart of Figure 18 there are three videos that stand out as severe outliers within the dataset: 13150_disgust, 11446_joy, and 2923_joy, with durations of 304.97, 235.07, and 41.04 seconds, respectively. Consequently, these videos also exhibit the highest frame counts.

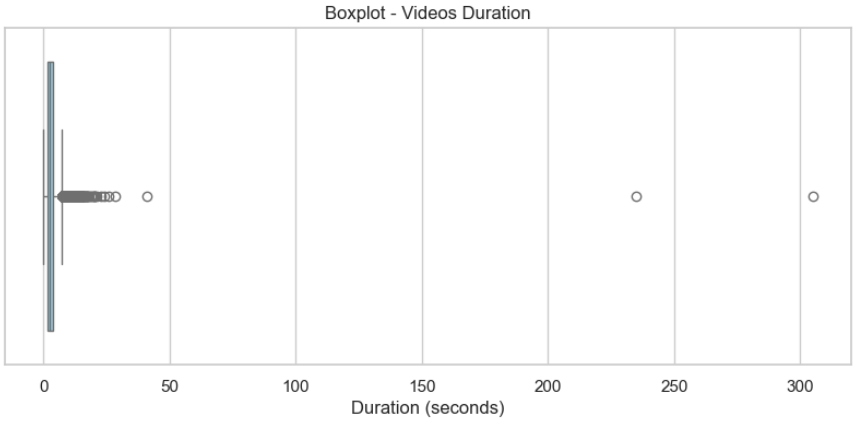


Figure 18 - Boxplot of the video durations in our dataset

6.1.2. Audio data understanding

After the removal of video 1165_neutral, we were left with a total of 13707 audio files. In terms of duration, there are no changes compared to the original videos. However, conducting a more in-depth analysis of potential audio characteristics remains a challenging task, primarily because these clips originate from television series, which often include background noise such as laughter and overlapping conversations. Nevertheless, according to the authors of the dataset, the clips were carefully selected to ensure that they contain only speech from a single speaker.

Regarding the sample rate, all audio files share a value of 44,100 Hz, which refers to the number of sound samples captured per second. This is the most commonly used sampling rate in digital audio, particularly in music and high-quality recordings. As for the number of frames, it ranges from values close to zero up to a maximum outlier of 13 449 177 frames. Naturally, this is directly associated with the duration of the audio clip. As shown in the of

Figure 19 the three extreme outlier values in the number of frames correspond directly to the same extreme values in the duration feature.

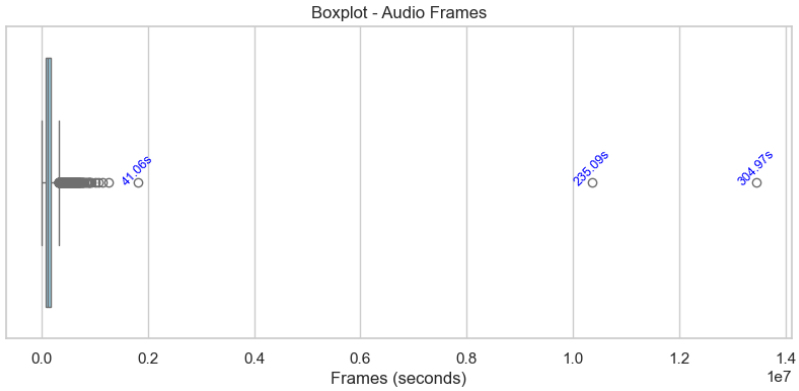


Figure 19 - Boxplot of audio frames in the dataset with annotations for the top 3 longest-duration clips

6.1.3. Text data understanding

It is important to note that, for this pre-processing step, we did not use the original .csv file associated with the dataset. Instead, we built a custom pipeline that processes the audio files and converts them into text. This approach was adopted in order to more closely simulate real-world conditions and ensure that the dataset reflects the actual challenges and characteristics of live data acquisition. Regarding the textual modality, our 13706 observations contain a total of 46168 unique words.

Concerning NER, it can be concluded from the table below that a significant portion of the dataset is composed of the entity types PERSON, CARDINAL, and DATE, which together account for approximately 75.4% of all observations. On the other hand, entities such as LAW, LANGUAGE, and EVENT are the least frequently observed, collectively representing only around 0.38% of the total occurrences.

Table 8 - NER from MELD text dataset without cleaning

NER	Count	%	NER	count	%
PERSON	2608	56.72	FAC	28	0.61
CARDINAL	485	10.55	PRODUCT	19	0.41
DATE	398	8.66	LOC	17	0.37
TIME	296	6.44	WORK_OF_ART	16	0.35
GPE	214	4.65	QUANTITY	12	0.26
ORG	190	4.13	PERCENT	12	0.26
ORDINAL	175	3.81	EVENT	8	0.17
NORP	65	1.41	LANGUAGE	3	0.07
MONEY	52	1.13			

It is also possible to analyse the sentiment within the text itself, particularly in terms of polarity and subjectivity, which present average values of approximately 0.166 and 0.546, respectively. A polarity value of 0.166 indicates a slightly positive sentiment, suggesting that, on average, the texts in the dataset tend to lean marginally toward a positive tone. The subjectivity score of 0.546 reflects a moderate level of subjectivity, implying that the texts generally contain personal opinions, emotions, or subjective expressions, rather than being purely factual or objective in nature. We can also examine the presence of empty values (i.e., samples without any text) within the dataset, as well as texts that contain two or less words, which in practice contribute little contextual information. The number of such cases was residual and is summarized in the table below.

Table 9 - Number of strings without text or with 2 or less words

Empty values	With two or less words	Total
217	20	237

At first glance, and without the need for extensive analysis, records containing empty strings in the text field are strong candidates for removal. These entries are likely to introduce noise into the emotion classification based on textual features and, consequently, would have a similar negative impact in a multimodal classification context.

6.1.4. Combining data understanding

Based on the previous data analysis, in addition to one video with a duration of zero minutes and six videos that were not referenced in the accompanying .csv files (and therefore were excluded), we currently identify a set of videos with atypical values, specifically, 831 outliers in terms of duration. Among these, three videos exhibit exceptionally high durations.

Including these videos in the final dataset could lead to significant conflicts during model training and evaluation. This is primarily due to the need for input sequences to be of uniform length, which would require applying padding or zero-filling to the shorter videos. The inclusion of extremely long clips would thus distort the distribution of sequence lengths and introduce excessive padding in a large portion of the dataset. As a consequence, the model could learn to overemphasize artificial patterns from the padding, leading to reduced generalization performance and potential bias toward longer sequences.

For these reasons, we considered the removal or separate treatment of these extreme cases to be a necessary step in maintaining the overall consistency and quality of the dataset used for modelling.

The remaining outliers were cross-referenced with the previous analyses, leading to the following conclusions: None of the outliers were related to observations where no textual content was present; The same applies to observations containing only one or two words; However, values with a resolution of 496x384 and 560x432 presented two and one instances, respectively, of videos without associated text.

Among all the values identified during the exploratory analysis, those related to non-standard resolution were deemed the least likely to negatively impact the model. Similarly, entries containing two or fewer words, in the context of the emotion analysis vector, were also considered to pose minimal constraint. Therefore, we decided to retain these values in the dataset.

Therefore, the rest of the nonstandard values, including outliers and missing textual content, resulted in 1048 instances considered as deviating from the expected pattern, with the distribution by emotion presented in the following table.

Table 10 - Pruning of the original dataset by excluding entries deemed inappropriate or unsuitable for analysis

Emotion	Count	% (aprox.)	Emotion	Count	% (aprox.)
Neutral	435	-7%	**Sadness**	114	-11%
Joy	202	-9%	**Disgust**	31	-9%
Anger	125	-8%	**Fear**	25	-7%
Suprisse	116	-7%	**Sadness**	114	-11%

On average, the dataset was reduced by approximately 8% of its original size. This reduction disproportionately affects intermediate emotions. However, since the removal of these values may enhance the clarity and quality of information processing by our model, we decided to proceed with their elimination.

6.2.Data Preparation

As previously discussed, we begin by cleaning the dataset, removing the entries identified and addressed in the preceding section. The following table summarizes the pre-processing steps applied to our dataset, along with a brief explanation of each operation performed.

Table 11 - Summary of the instances excluded from the dataset

Data	Reason
1165_neutral	0 seconds duration
13150_disgust, 11446_joy, 2923_joy	Extreme Outliers
IDs of empty lines (217 lines)	Without any valor for Emotion analysis vector
Outliers' values	Instances with durations classified as outliers based on statistical analysis

These data were removed, resulting in a dataset containing 12658 observations, with the distribution illustrated in the following chart.

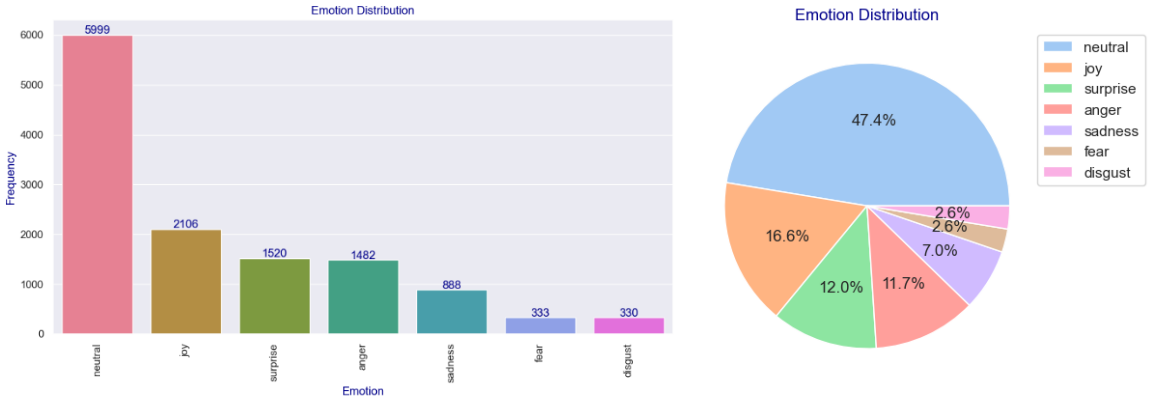


Figure 20 - Pie plot and Bar plot chart MELD after cleaning

With the dataset prepared for use, it was divided into training, validation, and test sets, with 60%, 20%, and 20% of the data allocated to each, respectively. The videos were distributed into separate folders to facilitate the subsequent creation of feature vectors based on their content[26]. The train, validation, and test splits were also created with consideration for the distribution of observations, ensuring that the proportion of emotions remained consistent across all three subsets, in alignment with the original dataset.

The same pipeline previously used during the data understanding phase was applied to separate the audio and, consequently, the text from the video. The approach used was identical to the one previously described. We used the large version of VideoMAE [80] or fine-tuning and video-to-vector transformation.

[26] The main idea was to use a function to split the videos into training, validation, and test sets. Then, for both the audio and text vectors, we used the same video split to ensure consistency across all modalities.

For the audio component, we selected Facebook's Wav2Vec2 XLSR-53 mode [81] originally designed for multilingual speech recognition, due to its strong performance in downstream tasks such as emotion classification.

Finally, for audio-to-text transcription, we used the large version of OpenAI's Whisper model, large-v3 [79]. For the emotion analysis on text, we used the large version of RoBERTa [36], for a powerful model to extract context from the dataset.

In the initial phase, prior to the construction of the multimodal model, we processed each modality independently, using the previously identified models. This allowed us to assess the capacity of each model to identify emotions within its respective vector representation.

6.3. Multimodal Fusion

Having reached the point where we address our three hypotheses, the approaches were designed based on three assumptions: For the first hypothesis, which relies on human knowledge in emotion classification, we apply a simple logic: in cases of conflicting emotional cues, the facial expression is considered the predominant emotion; For the second hypothesis, which adopts a unified perspective on emotions, we leverage transfer learning from previously studied models. Using an early fusion approach, we concatenate the emotion vectors from different modalities to perform the final classification; Finally, for the third hypothesis, we implement a late fusion strategy by applying a learned weighted combination of the emotion vectors. This assigns a relative importance to each modality in the final emotion prediction.

6.3.1. N-voting model

Conceptually, this approach assumes that the individual's true emotion within a video segment corresponds to the emotion most consistently detected across modalities. For example, if SER and text-based emotion analysis both identify fear, while FER detects happiness, the overall emotion assigned to the clip is fear. By default, in cases where all three modalities produce different emotion labels, the emotion predicted by FER is selected as the final classification.

The initial idea for the embryonic concept of multimodality is represented in *Multimodal Script 1st attempt.ipynb*, where, in a very summarized manner and as presented in the figure below, we approach the topic as follows:

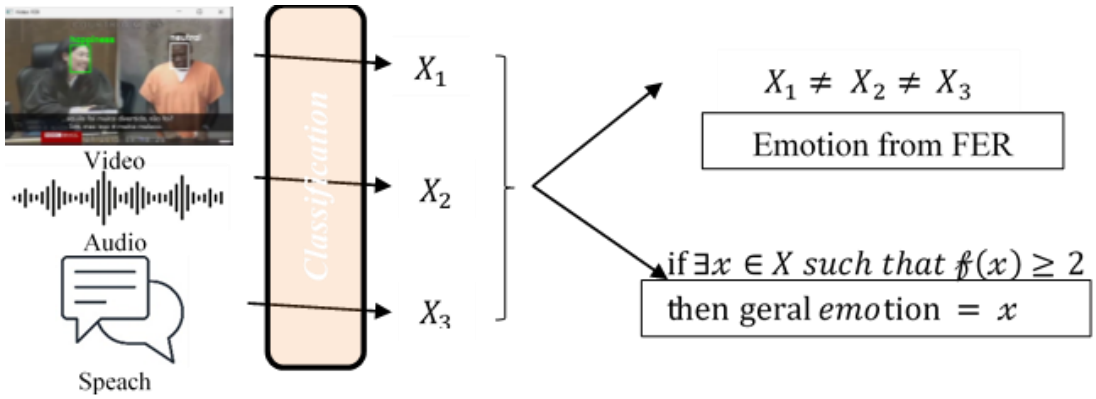


Figure 21 - Multimodal strategy H1

Video modelling

In order to create robust feature representations from video data, a systematic pipeline was implemented involving data augmentation, frame extraction, and feature extraction using a pretrained VideoMAE model.

Initially, a set of emotion-specific augmentation strategies was defined, categorized into basic, moderate, severe, and extreme levels, to enhance data diversity and mitigate overfitting, particularly for emotions with lower representation. The table below summarizes the data augmentation strategies implemented for each category.

Table 12 - Data augmentation for video

Extreme	Severe	Moderate	Basic
• Horizontal flip • Colors changes (ColorJitter[27]) • Affine[28] • Perspective • Erasing	• Rotation • Horizontal flip • Colors changes (ColorJitter) • Affine	• Rotation • Horizontal flip • Colors changes (ColorJitter)	None
Resize to 224*224[29]			

For each video sample, 16 frames were extracted in a temporally uniform manner, processed through the corresponding augmentation pipeline based on the target emotion, and prepared for model input. The VideoMAE feature extractor was then employed to process these

[27] Transform randomly changes the brightness, contrast, saturation, hue, and other properties of an image.
[28] Random affine transformation of the image keeping center invariant.
[29] Input to VideoMAE model.

frames, obtaining rich representations without retraining the base model. Instead of relying solely on the mean of the hidden states to summarize the features, a more informative approach was adopted by combining mean and max pooling over the hidden representations, capturing both average and most salient information across frames, as show at the next figure.

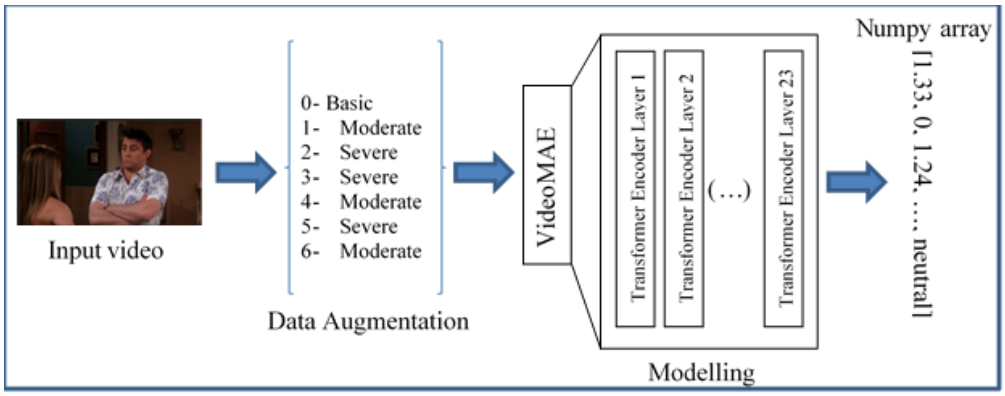


Figure 22 - Part A for the transfer learning approach - vector assembly

After feature extraction, the next step involved dimensionality reduction and classification training. Principal Component Analysis (PCA) was first applied to the extracted video features to reduce their dimensionality to 100 components. This step aimed to preserve the most informative aspects of the feature space while mitigating noise and redundancy, thus improving model efficiency and generalization. Subsequently, the dataset was split into training, validation, and test sets.

Figure 23 - t-SNE from video dataset (training Set)

The previous image represents and a t-SNE (t-Distributed Stochastic Neighbour Embedding) projection that was performed on the training set for visualization purposes. As we can see there are three major groups that can be identified, with one of them clearly divisible into

two subgroups. The blue cluster corresponds to the emotion neutral, which appears distinctly separated from the others. This isolation can be attributed to the large volume of data associated with this emotion.

The purple cluster represents the emotion joy, which, as a result of data augmentation and the use of class weighting, becomes distinguishable from the remaining emotions. The third group encompasses the rest of the emotional categories, which can be further subdivided into two subgroups: one dominated by sadness and anger, and another comprising surprise, fear, and disgust.

A simple feedforward neural network classifier was designed, composed of two linear layers with a ReLU activation and a dropout layer to reduce overfitting. The input layer matched the dimensionality of the PCA-reduced features. To address class imbalance, present in the dataset, balanced class weights were computed and incorporated into the loss function. Furthermore, label smoothing was applied to improve generalization and reduce overconfidence in the predictions. The classifier was trained using the Adam optimizer with weight decay for regularization and a cosine annealing learning rate scheduler to dynamically adjust the learning rate during training. The next image intent to show the second part of the transfer learning process.

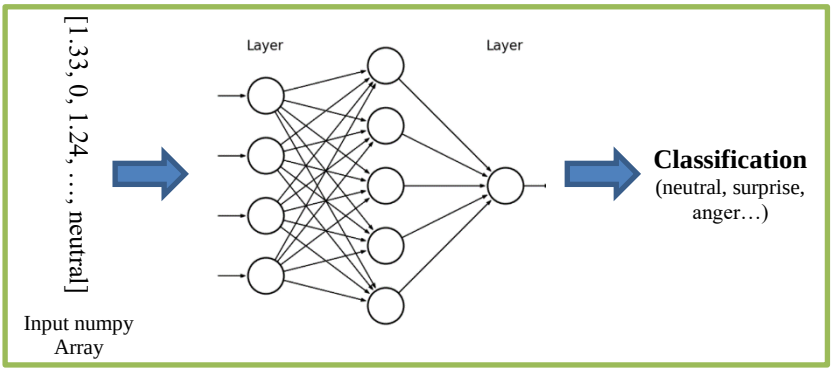


Figure 24 - Part B for the transfer learning approach - classification

Performance metrics, including loss and accuracy for both training and validation sets, were monitored across epochs, providing a comprehensive understanding of model behaviour during training.

Given the imbalance present in the dataset, the most appropriate evaluation metric is the F1-Score, as it better reflects the model's performance across all classes. Nonetheless, we also

report the Accuracy value, since the data pre-processing steps, including data augmentation, mitigate the impact of the imbalance, allowing this metric to also be considered meaningful.

Audio modelling

After organizing the dataset into training, validation, and test splits, the Wav2Vec2-Large XLSR-53 model was employed as a feature extractor for SER. To improve the model's robustness and its ability to generalize across variations in speech delivery, a data augmentation strategy was applied based on the emotion label. Four augmentation levels basic, moderate, severe, and extreme were defined, as referred at the next table. These included operations such as additive noise, pitch shifting, speed modification, and both time and frequency masking. The augmentation level was conditionally selected according to the emotion category, with more intense emotions (e.g., anger, fear, disgust) receiving heavier augmentation to simulate real-world variability.

Table 13 - Data augmentation for vector audio

Extreme	Severe	Moderate	Basic
• Strong Gaussian noise • Pitch shifting • Speed alteration • Time masking • Frequency masking	• Moderate Gaussian noise • Pitch shifting to alter voice frequency • Frequency masking, which occludes certain frequency bands	• Light Gaussian noise addition • Time masking, which occludes random time segments in the spectrogram	None

Each audio waveform was resampled to 16 kHz to match the model's expected input, followed by the corresponding augmentation. Features were extracted by passing the processed waveform through the Wav2Vec2 model. Instead of relying solely on the average of the hidden states, a combined strategy using both mean and max pooling was applied across the last hidden layer to generate a richer and more discriminative feature representation.

The same approach applied to the video dataset was adopted for the audio dataset, using Wav2Vec2 to extract the last hidden layer, which was then passed through a classifier to generate logits and produce the final emotion classification. The extracted features were flattened and then subjected to PCA the result is at the next figure. It is difficult for the model to distinguish clear emotion groups, as the applied pre-processing results in a mixing of almost all emotions.

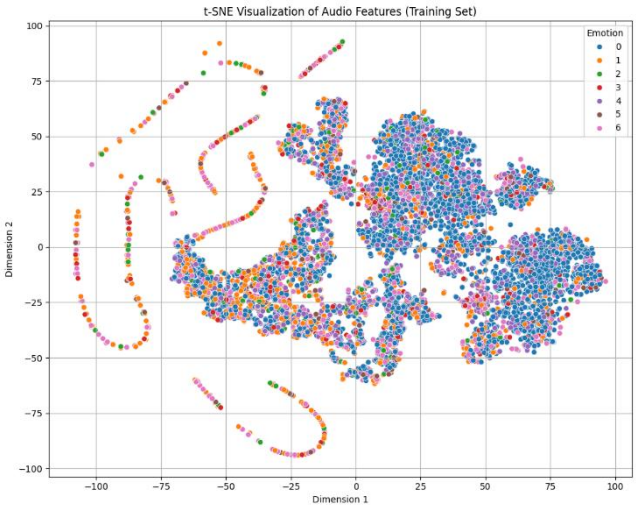


Figure 25 - t-SNE from audio dataset (training Set)

The model was trained using class-balanced weights and label smoothing, with the Adam optimizer and a cosine annealing learning rate scheduler. Early stopping was applied based on the validation weighted F1-score, preserving the best model and avoiding overfitting.

We can state that, on its own, the audio analysis of the dataset does not significantly contribute to the classification of emotions.

Text modelling

To address the emotion recognition task from textual input, a text pre-processing and semantic enrichment pipeline was then applied. This pipeline began with normalization steps including lowercasing, removal of user handles, URLs, punctuation, and stopwords using the neattext library[30]. NER, POS tagging, and lemmatization were performed using the spaCy large English model. Additionally, each utterance was scored using VADER for sentiment polarity.

To improve quality and reduce noise, the cleaned texts were processed with a TF-IDF-based filtering strategy, retaining only terms with importance above a defined threshold. Final cleaning steps removed remaining digits and punctuation, followed by tokenization with NLTK.

[30] Available at https://pypi.org/project/neattext/

To increase data diversity and help the model generalize, a conditional text data augmentation strategy was introduced using EDA (Easy Data Augmentation). Each emotion label was mapped to an augmentation level basic, moderate, severe, or extreme which determined the number of EDA operations applied. Minority or high-variability classes such as fear, disgust, and anger received stronger augmentation.

After pre-processing and augmentation, text inputs were encoded using the pretrained RoBERTa-large model. Each utterance was tokenized and passed through the model, and a combined representation was obtained by concatenating mean and max pooling over the last hidden state. Subsequently, dimensionality reduction was applied using PCA, reducing the embedding size to 100 dimensions. The t-SNE chart originated was the follow:

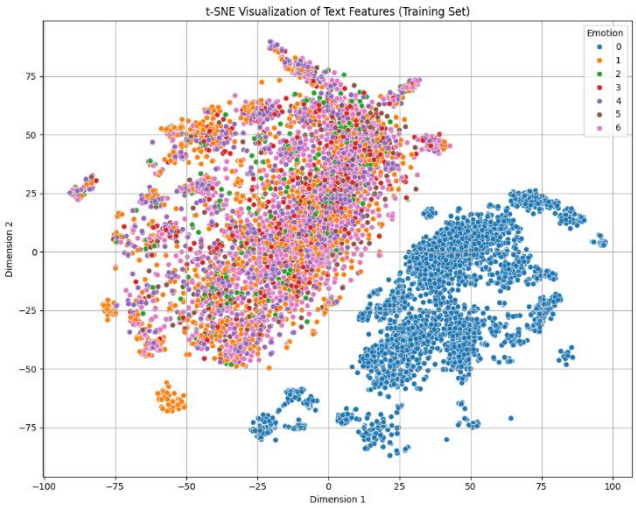


Figure 26 - t-SNE from text dataset (training Set)

As observed in the previous plot, two major clusters can be clearly identified: one predominantly composed of the neutral emotion represented by the blue color, and another grouping the remaining emotions.

For the classification task, we employed the same approach applied to the SER and FER feature vectors. The model was trained using class-balanced weights and label smoothing to address class imbalance and encourage generalization. The training loop used the Adam optimizer with weight decay, and performance was evaluated on the validation set using the weighted F1-score. Early stopping was implemented based on stagnation in validation performance.

6.3.2. Early fusion model

Each video file was first processed to extract a fixed number of 16 evenly spaced frames. These frames were resized to 224×224 pixels and converted into RGB tensors. Once the frames were pre-processed, they were passed through the VideoMAE large model to obtain the final hidden states. Two types of temporal pooling strategies mean pooling and max pooling were applied across the sequence of frame embeddings. These pooled representations were concatenated, resulting in a final 2048-dimensional feature vector per video sample. The features and corresponding emotion labels were saved separately for the training, validation, and test sets under the label FER.

The audio stream was processed using the Wav2Vec2-Large XLSR-53 model. Each audio clip, originally extracted from the videos, was first validated to ensure a mono format, converting stereo files by averaging the channels. Additionally, all samples were resampled to a 16 kHz sampling rate to match the model's requirements. The resampled waveforms were passed through the Wav2Vec2 model, from which the final hidden states were extracted. Similar to the video stream, mean pooling and max pooling were computed across the time dimension and concatenated to obtain a 2048-dimensional embedding for each audio clip. These embeddings were stored as the SER feature representations.

Textual features were derived from the transcriptions of the audio data. The inputs were passed through a pretrained RoBERTa-large model, and both mean and max pooling were applied across the token embeddings to generate a robust 2048-dimensional representation.

In the early fusion approach, the core idea was to concatenate all feature vectors into a single unified vector. Although alternative vector combination techniques could have been employed, the standard concatenation method was chosen for simplicity and consistency, given that this is a proof of concept. The following figure illustrates the architecture adopted in this approach[31].

[31] Note: Only the MLP with ReLU architecture is illustrated in the figure; however, the experiments were conducted using four different classifiers: a simple MLP with ReLU, an MLP combined with an LSTM module, an MLP with a Transformer module, and an MLP with an attention mechanism.

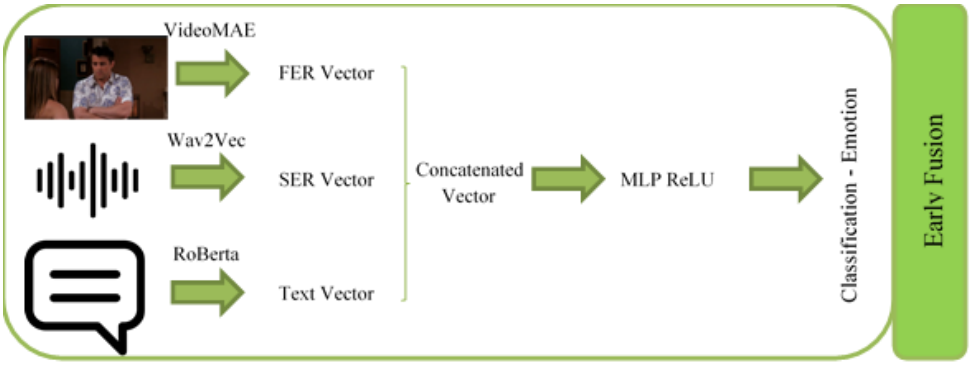


Figure 27 - Multimodal approach, early fusion architecture.

Once the unified vector was formed, it was passed through a classifier. Additionally, instead of limiting the evaluation to a single simple classifier, we tested three different types of classifiers: a basic MLP, an MLP augmented with an attention mechanism (MLP-A), and finally, an LSTM followed by an MLP (LSTM + MLP). This experimental design also allowed us to assess which neural network architecture was best suited for classifying the emotion vectors.

Using the optuna library [82], we found the best values for our training model: learning rate and weight decay which fits better in our problem[32].

In addition to testing multiple classifiers, we applied four distinct data preparation strategies during the training phase of our final classifier. The first approach consisted of a basic pipeline without data augmentation and without dimensionality reduction. The second approach included both data augmentation and dimensionality reduction. The third strategy involved training without data augmentation but with an increased dataset size. Finally, the fourth and most comprehensive approach combined data augmentation with an increased dataset. These configurations were implemented to evaluate how different preparation strategies impact classification performance, especially in the context of multimodal fusion.

[32] (Learning rate: 0.0005238452508510812, the weight decay: 2.331735062538531e-05)

6.3.3. Late fusion model

This process is ultimately more streamlined, the only requirement at this stage is to define the necessary pipelines to extract the output from the fully connected layers of each modality's MLP-based classifier and convert it into a probability vector using the SoftMax function. Once these probability vectors are obtained, the final step involves determining the values of K_1, K_2 *and* K_3 as illustrated in the next figure.

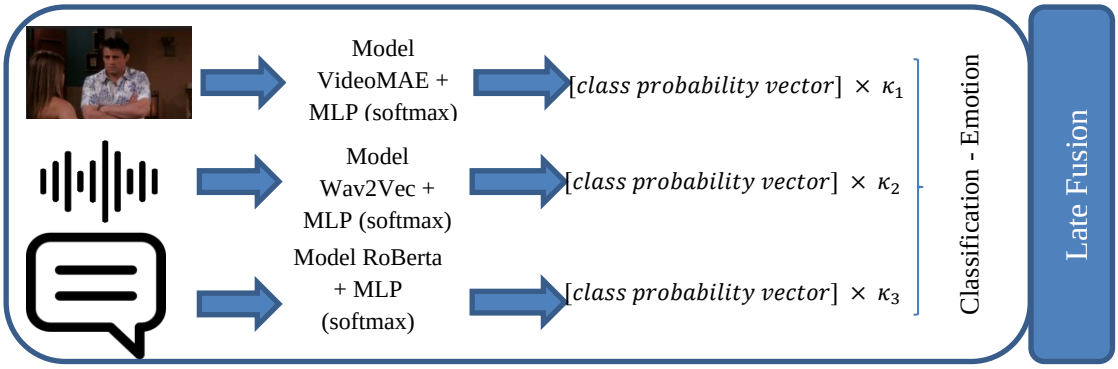


Figure 28 - Multimodal approach, late fusion architecture

In this framework, modality-specific feature vectors extracted independently by VideoMAE for facial expressions, Wav2Vec for speech, and RoBERTa for text are processed in parallel. Each vector is passed through a two-layer MLP that produces a probability distribution over the target emotion categories. The values of K_1, K_2, and K_3 were learned using a simple neural network, which essentially aimed to determine the optimal weight values (K) for our classification problem.

6.4. Testing Multimodal model form emotion analysis

Within the multimodal approach, the process encompassed the evaluation of a range of models, beginning with the individual analysis of each vector and culminating in a joint multimodal strategy, where three hypotheses were formulated to address the present problem. Regarding the results obtained from the unimodal analysis, it can be concluded that video exhibits a certain predominance over audio and text when evaluated independently. This finding lends support to the hypothesis that, in cases of conflict between modalities, facial emotion may carry greater weight and relevance in determining the overall emotional state.

Table 14 - Single vector test metrics with data augmentation and with data augmentation and dataset increased

	With Data Augmentation		With Data Augmentation and Dataset increased	
	F1	**ACC**	**F1**	**ACC**
Video - VideoMAE	71.10	69.17	**74.40**	**72.46**
Audio – Wav2vec	30.81	32.30	42.23	42.88
Text - Roberta	59.54	58.01	58.92	57.70

Applied to the pipeline, the N-voting model was used to identify the most frequently detected emotion across the three modalities. In cases of disagreement, the emotion predicted by the FER component was given precedence as the final classification. The results obtained showed an accuracy of 50.50% and an F1-score of 52.50%.

Regarding our second hypothesis (H2), corresponding to the Early Fusion strategy, we explored the impact of passing the last hidden state representations through three different classifiers: MLP, MLP-A, and LSTM combined with MLP. Furthermore, we examined how these classifiers performed under different dataset configurations, including the absence or presence of data augmentation, the application of majority class balancing through dimensionality reduction, and the use of an expanded dataset. In total, four experimental conditions were considered: without data augmentation (DA) and with majority class balancing using dimensionality reduction (DR); with data augmentation and majority class balancing using dimensionality reduction; without data augmentation and with an expanded dataset; and finally, with both data augmentation and dataset expansion.

Table 15 - Summary table of the metrics obtained from the Early Fusion model representing Hypothesis 2.

	MLP		MLP-A		LSTM + MLP	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**
Without DA and DR	45.87	44.87	3.99	17.99	49.02	50.07
With DA and DR	**50.89**	**50.54**	33.15	32.60	43.37	43.20
With DA and with Dataset increased	44.87	45.13	8.99	13.99	44.98	42.99
Without DA and with Dataset increased	48.34	47.93	11.91	14.11	48.63	48.44

The late fusion process consists of taking the NumPy arrays obtained from the hidden connected layers of each modality and passing them through a final layer that produces a probability vector. This vector is then fed into a trainable weighted ensemble model that assigns a learnable weight to each modality (visual, audio, and text). The weights are normalized using a softmax function and the final output is a weighted combination of the individual probability vectors.

The weights are optimized during training to maximize the overall performance of the ensemble. Given the imbalanced nature of our dataset, the F1 score was used as the evaluation metric to guide this optimization, as it provides a more balanced assessment of classification performance across classes.

Table 16 - Summary table of the approaches followed in the multimodal pipeline

Hypotheses model	ACC	F1
N-voting (H1)	50.50	52.50
Early Fusion (H2)	50.89	50.54
Late Fusion (H3)[33]	**53.05**	**58.07**

The results demonstrate that the Late Fusion strategy, which relies on learned weights to combine modalities, proved to be the most efficient and reliable method for multimodal emotion recognition within the scope of this research.

6.5. Summary

This chapter focuses on the development of a unified emotion evaluation framework capable of managing potential inconsistencies between the emotions detected by different modalities. The approach explores the integration of FER, SER, and emotion analysis from text within a multimodal emotion recognition system. Due to the lack of publicly available datasets involving criminal interrogations, the MELD was selected as the experimental basis, as it approximates real-world interactive contexts.

The work conducted in this chapter investigated three main hypotheses regarding the relative contribution of each modality to emotion recognition: first, that facial emotion should dominate in cases of conflict between modalities (N-voting strategy); second, that all modalities contribute equally to a unified emotion assessment (Early Fusion strategy); and third, that each modality contributes differently depending on its relevance, with weights learned automatically by the model (Late Fusion strategy).

Features were extracted from each modality using pre-trained models adapted for this task: VideoMAE Large for video, Wav2Vec2 XLSR-53 Large for audio, and RoBERTa Large for text. To summarise the extracted features, pooling techniques (mean and max) and

[33] Final weights: K=0.077, L=0.075, M=0.847

dimensionality reduction via PCA were applied. An initial unimodal analysis was carried out for each vector individually.

Following this, the three fusion strategies corresponding to the proposed hypotheses were implemented. The N-voting strategy applied facial dominance in the final decision-making. Early Fusion involved the concatenation of feature vectors from each modality before classification, with multiple classifiers tested, including MLP and LSTM combined with MLP. Late Fusion combined the output probabilities from each unimodal classifier using learned weights, allowing the model to optimise the contribution of each modality in producing the final emotion classification.

The multimodal evaluation revealed that, among the unimodal approaches, video data outperformed both audio and text, suggesting that facial expressions hold greater significance in isolated emotion recognition tasks. This supports the Hypothesis 1 where facial cues dominate in cases of inter-modal disagreement, achieving 52.50% F1-score. Hypothesis 2, which involved concatenating hidden states from all modalities and evaluating multiple classifiers (MLP, MLP-A, and LSTM+MLP), yielded mixed results across data augmentation and balancing configurations, with best performance reaching 50.54% F1-score. For last, Hypothesis 3, which employed a trainable weighted ensemble to adaptively combine modality-specific predictions, delivered the most promising results, achieving 58.07% F1-score and 53.05% accuracy.

7. Conclusion and Future Work

This dissertation presents INTU-AI, a prototype system developed to support the PJM in conducting and documenting police interrogations through multimodal emotion recognition. The implementation of this tool represents a significant step towards the integration of AI into traditionally human-driven processes.

INTU-AI was conceived and developed as a fully integrated, Python-based end-to-end system capable of processing video, audio, and real-time camera input. The system segments the incoming data into 10-second intervals, during which it applies FER, SER, and text-based emotion analysis. The system outputs include a complete transcript of the interrogation, a visual annotation of the detected emotions across modalities, and an automatically generated official PJM report.

A core challenge addressed in this thesis was the reconciliation of inconsistent emotion predictions across modalities. To tackle this, three fusion strategies were explored using the MELD dataset, which closely reflects natural, real-world conversational dynamics. The strategies included: (1) a majority voting mechanism giving priority to facial expressions (N-Voting); (2) an early fusion strategy treating all modalities equally by concatenating their respective feature vectors; and (3) a late fusion method, in which the model learns to assign optimal weights to each modality during training.

Among these strategies, the late fusion approach delivered the best performance, achieving an accuracy of 53.05% and a macro F1-score of 58.07%, the results demonstrate that allowing the model to learn the relative importance of each modality yields more robust emotion classification outcomes. As such, INTU-AI provides a promising proof of concept for the application of AI in forensic settings and lays the groundwork for the development of more advanced and operationally integrated systems for emotion analysis in criminal investigations.

The future work involves adapting the machine learning pipeline to real-world interrogation datasets. While the current system was validated on publicly available resources such as MELD and AffectAlchemy, these datasets lack the contextual specificity and emotional nuance present in actual police interrogations. Access to carefully annotated, domain-specific data would significantly improve the model's ability to generalize and make accurate inferences in realistic scenarios.

8. Bibliography

[1] M. Suleyman, The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma, New York: Crown, 2023.

[2] J. Garcia, "Multimodal Approach," 2025. [Online]. Available: https://drive.google.com/drive/folders/1yMLaJfmgDQkcrA-uP6X4oBKqkhga2oRJ. [Accessed 01 May 2025].

[3] J. Garcia, "INTU-IA program," 2025. [Online]. Available: https://drive.google.com/drive/folders/1U1xONIXLtSniIxZP386VrekKvyF0x-FZ. [Accessed 01 May 2025].

[4] T. Kanade, "Picture Processing System by Computer Complex and Recognition of Human Faces," Ph.D. dissertation, Department of Information Science, Kyoto University, Kyoto, Japan, 1973.

[5] A. Pentland and T. Choudhury, "Personalizing Smart Environments: Face Recognition for Human Interaction," *IEEE Computer,* vol. 33, no. 2, p. 50–55, 2000.

[6] Gaya-Morey and F. Xavier, "Unveiling the Human-like Similarities of Automatic Facial Expression Recognition: An Empirical Exploration through Explainable AI," *Multimedia Tools and Applications,* vol. 83, no. 38, p. 85725–85753, 2024.

[7] Ş. İ. Serengil and A. Özpinar, "A Benchmark of Facial Recognition Pipelines," *Gazi University Journal of Information Technologies,* pp. 1-14, 29 March 2023.

[8] M. Jin, "A Study of Face Alignment Methods in Unmasked and Masked Face Recognition," (M.S. Dissertation, Department of Information Technology, Uppsala University, Sweden, 2023.

[9] K. Vemou, A. Horvath and T. Zerdick, "Facial Emotion Recognition," *TechDispatch,* no. 1, pp. 1-5, 2021.

[10] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias and W. Fellenz, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine,* vol. 18, no. 1, p. 32–80, January January 2001.

[11] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art," 30 March 2012. [Online]. Available: https://arxiv.org/pdf/1203.6722. [Accessed 14 11 2024].

[12] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions On Affective Computing,* vol. 13, no. 3, pp. 1195 - 1215, 2022.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou and B. Schuller, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in *2016 IEEE ICASSP in Shanghai*, Shanghai, China, 2016.

[14] L. Pepino, “Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings,” in *Interspeech*, Brno, Czech Republic, 2021.

[15] F. Barbieri, J. Collados, L. Neves and L. Anke, “TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification,” *EMNLP*, p. 1644–1650, November 2020.

[16] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, “CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset,” *IEEE Transactions on Affective Computing,* vol. 5, no. 6, p. 377–390, 2014.

[17] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLOS ONE,* vol. 13, no. 5, 2018.

[18] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee and S. Narayanan, “IEMOCAP: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[19] S. Poria, E. Cambria, R. Bajpai and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, p. 98–125, 2017.

[20] M. Shkhanukova, “ASR state-of-the-art: Wav2Vec, Whisper, DeepSpeech,” Medium, 4 December 2022. [Online]. Available: https://medium.com/@milana.shxanukova15/asr-state-of-the-art-wav2vec-whisper-deepspeech-e1b715c2aed0. [Accessed 13 January 2025].

[21] OpenAI, “Whisper: Open-source speech recognition,” OpenAI, 21 September 2022. [Online]. Available: https://openai.com/index/whisper/. [Accessed 12 December 2024].

[22] L. Roberts, “Understanding the Mel Spectrogram,” 06 March 2020. [Online]. Available: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53. [Accessed 28 Decemebr 2024].

[23] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” in *EMNLP 2004, Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

[24] Y. Liu and M. Lapata, “Text Summarization with Pretrained Encoders,” in *EMNLP-IJCNLP 2019*, Hong Kong, China, 2019.

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” in *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, 2020.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. Liu, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research,* vol. 21, pp. 1-67, 2020.

[27] AWS, “What is sentiment analysis?,” [Online]. Available: https://aws.amazon.com/what-is/sentiment-analysis/. [Accessed 2025 January 13].

[28] R. 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,* vol. 89, no. 4, p. 344–350, 2001.

[29] P. Ekman, “Basic emotions,” in *Handbook of Cognition and Emotion,* edited by Tim Dalgleish and Mick J. Power, published by Wiley in 1999, pages 45–6, 1999.

[30] C. Strapparava and A. Valitutti, “WordNet Affect: an Affective Extension of WordNet,” in *LREC 2004*, Lisbon, Portugal, 2004.

[31] A. Esuli and F. Sebastiani, “SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining,” in *LREC 2006, Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006.

[32] B. Pang, L. Lee and S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in *EMNLP 2002*, Philadelphia, 2002.

[33] S. M. Mohammad and P. D. Turney, “Crowdsourcing a Word-Emotion Association Lexicon,” *Computational Intelligence,* vol. 29, no. 3, p. 436–465, 2013.

[34] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng and C. Potts, “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,” *EMNLP,* p. 1631–1642, Octuber 2013.

[35] J. Devlin, M. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL-HLT*, Minneapolis, 2019.

[36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *Proc. 8th Int. Conf. Learn. Representations (ICLR),* 2020.

[37] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh and L. Morency, “Context-Dependent Sentiment Analysis in User-Generated Videos,” *ACL,* vol. 1, p. 873–883, 2017.

[38] S. Ji, W. Xu, M. Yang and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition,” in *ICML*, Haifa, Israel, 2010.

[39] D. Kollias, P. Tzirakis, M. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia and S. Zafeiriou, “Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond,” *International Journal of Computer Vision*, vol. 127, no. 6–7, p. 907–929, February 2019.

[40] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić and C. Schmid, “ViViT: A Video Vision Transformer,” *ICCV,* p. 6836–6846, October 2021.

[41] G. Bertasius, H. Wang and L. Torresani, “Is Space-Time Attention All You Need for Video Understanding?,” *38th International Conference on Machine Learning*, p. 9982–9992, 2021.

[42] T. Baltrušaitis, P. Robinson and L. Morency, “OpenFace: an open source facial behavior analysis toolkit,” *IEEE WACV,* p. 1–10, March 2016.

[43] FEW Consortium, “Acted Facial Expressions In The Wild,” 1 February 2013. [Online]. Available: https://users.cecs.anu.edu.au/~few_group/AFEW.html. [Accessed 16 January 2025].

[44] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria and R. Mihalcea, “MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations,” in *ACL 2019, proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence*, Italy, July 2019.

[45] A. Baevski, H. Zhou, A. Mohamed and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Advances in Neural Information Processing Systems*, December, 2020.

[46] W. Hsu, B. Bolte, Y. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed, “HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, April 2021.

[47] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, Z. Wei, Y. Qian, J. Li and F. Wei, "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing," *ACL 2022 (60th Annual Meeting)*, p. 5723–5738, May 2021.

[48] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, p. 1509–1525, 2021.

[49] A. Singh, "Mastering Sliding Window Techniques," Medium, 08 August 2023. [Online]. Available: https://medium.com/@rishu__2701/mastering-sliding-window-techniques-48f819194fd7. [Accessed 2025 January 17].

[50] M. Stent, "Dynamic Time Warping," Medium, 9 April 2024. [Online]. Available: https://medium.com/@markstent/dynamic-time-warping-a8c5027defb6. [Accessed 2025 January 20].

[51] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," Workshop on Energy Efficient Machine Learning and Cognitive Computing, 2019.

[52] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh and Dario, "Language Models are Few-Shot Learners," *NeurIPS 2020*, vol. 33, p. 1877–1901, December 2020.

[53] Open AI, "GPT-4 Technical Report," Open AI, 4 March 2024. [Online]. Available: https://arxiv.org/pdf/2303.08774. [Accessed 7 November 2024].

[54] D. M. Melanchthon, "CMU-MOSEI Dataset," *ACL*, vol. 1, p. 2236–2246, September 2018.

[55] K. Gadzicki, R. Khamsehashari and C. Zetzsche, "Early vs Late Fusion in Multimodal Convolutional Neural Networks," *FUSION*, p. 1–6, 2020.

[56] Z. Lian, H. Sun, L. Sun, H. Chen, L. Chen, H. Gu, Z. Wen, S. Chen, S. Zhang, H. Yao, B. Liu, R. Liu, S. Liang, Y. Li, J. Yi and J. Tao, "OV-MER: Towards Open-Vocabulary Multimodal Emotion Recognition," Conference on Computer Vision and Pattern Recognition, 2025.

[57] Z. Lian, H. Sun, L. Sun, Z. Wen, S. Zhang, S. Chen, H. Gu, J. Zhao, Z. Ma, X. Chen, J. Yi, R. Liu, K. Xu, B. Liu, E. Cambria, G. Zhao, B. W. Schuller and J. Tao, "MER 2024: Semi-Supervised Learning, Noise Robustness, and Open-Vocabulary

Multimodal Emotion Recognition," *Proc. 2nd Int. Workshop on Multimodal and Responsible Affective Computing,* p. 41–48, 1 November 2024.

[58] P. Durai, "Ghithub," 2023. [Online]. Available: https://github.com/spmallick/learnopencv/tree/master/Facial-Emotion-Recognition. [Accessed 10 November 2024].

[59] GeeksforGeeks, "GG Net Architecture Explained," 07 june 2024. [Online]. Available: https://www.geeksforgeeks.org/vgg-net-architecture-explained/. [Accessed 27 january 2025].

[60] Restack, "Vgg Face Model Architecture Overview," 05 may 2024. [Online]. Available: https://www.restack.io/p/vgg-face-model-answer-ai-synthesis-case-studies-cat-ai. [Accessed 28 January 2025].

[61] Linzaer, "Ultra Light Fast Generic Face Detector," 2019. [Online]. Available: https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB. [Accessed 27 January 2025].

[62] J. Grosman, "Hugging Face," 2021. [Online]. Available: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english. [Accessed 10 January 2025].

[63] J. Grosman, "Dataloop," [Online]. Available: https://dataloop.ai/library/model/jonatasgrosman_wav2vec2-large-xlsr-53-english/. [Accessed 10 January 2025].

[64] A. Kapase, N. Uke, J. Savant, M. Desai, S. Ghatage and A. Rahangdal, ""AffectAlchemy": An Affective Dataset Based on Plutchik's Psychological Model for Text-Based Emotion Recognition and its Analysis Using ML Techniques.," IEEE, 2024, Pune, India, 2024.

[65] R. Plutchik, Emotion: A Psychoevolutionary Synthesis., USA: Harper & Row, 1980.

[66] A. Kapase, "affect-alchemy," 24 August 2024. [Online]. Available: https://github.com/ajaykapase/affect-alchemy. [Accessed 22 January 2025].

[67] A. Zhang, Z. Lipton, M. Li and A. Smola, "Dive into Deep Learning," 2019. [Online]. Available: https://d2l.ai. [Accessed 30 January 2025].

[68] J. Li, X. Wang and Z. Zeng, "Tracing Intricate Cues in Dialogue: Joint Graph Structure and Sentiment Dynamics for Multimodal Emotion Recognition," 60th Annual Meeting of the Association for Computational Linguistics, 2024.

[69] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, "CRISP-DM 1.0," NCR Systems Engineering Copenhagen, 2000.

[70] S. Pragnya, "VADER (Valence Aware Dictionary and sentiment Reasoner) Sentiment Analysis," Medium, 16 January 2022. [Online]. Available: https://swayanshu.medium.com/vader-valence-aware-dictionary-and-sentiment-reasoner-sentiment-analysis-28251536698. [Accessed 13 March 2025].

[71] Invideo, [Online]. Available: https://invideo.io/. [Accessed 26 December 2024].

[72] P. Ekman and W. Friesen, "Constants across Cultures in the Face and Emotion," *Journal of Personality and Social Psychology,* pp. 124-129, 1971.

[73] P. Ekman, "An argument for basic emotions," *Cognition and Emotion,* vol. 6, no. 3-4, p. 169–200, 08 October 1992.

[74] C. E. Izard, "Innate and universal facial expressions: Evidence from developmental and cross-cultural research. Psychological Bulletin," *Psychological Bulletin,* vol. 115, no. 2, pp. 288-299, April 1994.

[75] L. J. Keltner D, "Emotion," in *Handbook of Social Psychology,* New Jersey, USA: Wiley, 2010, pp. 317-352.

[76] C. Darwin, The Expression of the Emotions in Man and Animals., London: John Murray, 1872.

[77] S. Poria, "MELD: Multimodal EmotionLines Dataset," 2018. [Online]. Available: https://affective-meld.github.io/. [Accessed 1 January 2025].

[78] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *In Proceedings of the 40th International Conference on Machine Learning,* vol. Vol. 202., p. 28492–28518, 2023.

[79] Z. Tong, Y. Song, J. Wang and L. Wang, "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," in *NeurIPS 2022,* March 2022.

[80] Q. Xu, A. Baevski and M. Auli, "Simple and Effective Zero-shot Cross-lingual Phoneme Recognition," Proceedings of Interspeech 2018, Hyderabad India, 2021.

[81] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* Anchorage, Alaska USA, 2019.

[82] S. Bertrand, “RetinaFace,” 23 December 2024. [Online]. Available: https://github.com/serengil/retinaface. [Accessed 13 January 2025].

[83] I. Centeno, “MTCNN,” 23 Dezember 2024. [Online]. Available: https://github.com/ipazc/mtcnn. [Accessed 10 February 2024].

[84] Facebook Open Source, “FastText,” Facebook, 2022. [Online]. Available: https://fasttext.cc/. [Accessed 12 January 2025].

[85] X. Tang, Y. Lin, T. Dang, Y. Zhang and J. Cheng, “Speech Emotion Recognition Via CNN-Transformer and Multidimensional Attention Mechanism,” *Speech Communication,* vol. 171, no. 103242, 2025.

[86] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 4690-4699, 2019.

[87] P. Viola and M. Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features,” CVPR 2001, Kauai, Hawaii, 2001.

[88] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. C. Berg, “SSD: Single Shot MultiBox Detector,” *ECCV,* vol. 9905, pp. 21-37, 2016.

[89] O. Parkhi, “Deep Face Recognition,” BMVC, Swansea UK, 2015.

[90] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, “Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks,” *IEEE signal processing letters,* vol. 23, no. 10, pp. 1499-1503, 2016.

10.1. ANNEX A- INTU-IA - USER GUIDE - VERSION 1.0

User Manual

Introduction

This project serves as a support tool for criminal investigations, assisting investigators in the interrogation of witnesses and suspects.
Developed as a modular solution, the INTU-AI program is an innovative system capable of categorizing human emotions through three distinct approaches:

- FER (Facial Emotion Recognition) – Identifies emotions through facial expressions.
- SER (Speech Emotion Recognition) – Analyses vocal tone to detect emotional states.
- Sentiment Analysis – Classifies the sentiment expressed in spoken words.

Beyond being a powerful tool for enhancing interrogations, INTU-AI fully automates the administrative process associated with investigations. The system extracts subject data directly from official records or national identification cards, enabling the automatic completion of standard reports used by the Military Judiciary Police (PJM).
Additionally, leveraging Natural Language Processing (NLP) techniques, the program transcribes entire interrogations from audio to text, providing investigators with instant access to full transcripts.
As a final output, the investigator receives:

- Official PJM reports, automatically populated with relevant data.
- A complementary report detailing emotion analysis throughout the interrogation, recording emotional fluctuations every 10 seconds.
- A video overlay combining insights from all three emotional analysis methods (FER, SER, and sentiment analysis).

This tool represents a significant advancement in criminal investigation, improving efficiency, accuracy, and analytical support during interrogations.

Technical Requirements

Machine Requirements:

Requirement	Minimum	Preferred
Camera	USB-connected camera34	
Microphone	USB-connected microphone35	
RAM	8 gb	16 GiB
Disk Space	5gb36	
Operating System	Windows XP	Windows 10 or later
Supported Video Formats	*.mp4;*.avi;*.mov	
Supported Audio Formats	*.mp3;*.wav;*.ogg”	

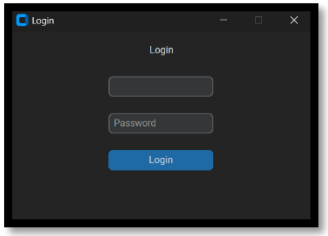
[34] For the streaming functionality, the program operates independently for all other features.
[35] The same that 1.

[36] The program follows a logic where all information is stored within the installation directory. Each time the program analyses an interview, it creates a dedicated folder within the main directory to store all related data. Consequently, this will inherently require additional storage space.

Interview Requirements

The video or live stream should include a maximum of three people.
For optimal facial feature capture, the interviewee should be centred on the screen and positioned at an appropriate distance, occupying at least one-third of the image.
The background of the recording should be as neutral and simple as possible, avoiding posters, photographs, or any visual elements that could be mistakenly identified by the program as faces.

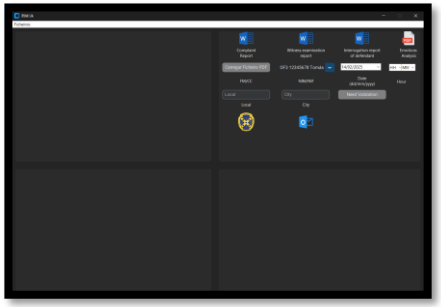
Operating Module



The program starts with an authentication window, requiring a username and password.
For testing purposes, the credentials are:
Username – "kl3z"
Password – "12345"
The main window is divided into four distinct sections, each serving a specific function in presenting processed information:

1. Top-left section: Displays the interrogated individual's data.
2. Bottom-left section: Presents a graphical representation of Facial Emotion Recognition (FER), showing the count of detected emotions.
3. Bottom-right section: Displays the Speech Emotion Recognition (SER) graph, tracking the frequency of identified emotions.
4. Top-right section: Functions as a control panel, allowing document selection and the input of interview-related information.

This structured layout ensures an intuitive and efficient workflow for investigators.



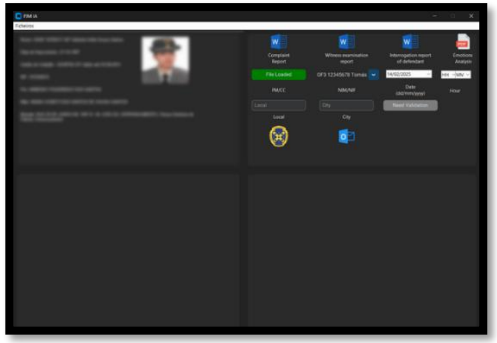
In the top-right corner, several options are available, including three Word documents:

- "Complaint Report"
- "Witness Examination Report"
- "Interrogation Report of Defendant"

These documents are officially used by the PJM, each corresponding to a specific type of case.
A PDF document is also available, containing an initial section that includes details about the interviewee, interviewer, and interview location, along with an introductory summary of the session (limited to 1600 characters).
Within the same document:

- Annex A contains the full transcription of the interrogation.
- Subsequent annexes present the sentiment analysis results, structured in 10-second intervals, displaying the emotion classification for each analysed vector (FER, SER, and text sentiment analysis).

Additionally, on the second row of options, there is a button labeled “Load PDF File”, designed to import the interviewee’s MF (Folha de Matrícula) or CC (Cartão de Cidadão). Once the information is successfully uploaded, the button will turn green, indicating completion.



On the right side of the previously mentioned button, there is a combo box displaying the NIM and the name of the interrogators[37]. Next to it, there is another combo box for selecting the interview date, which must be filled in according to the provided interview information.
Similarly, the time field is divided into two separate combo boxes, allowing the user to specify the hour and minutes of the interview.
The second-to-last row of menu options includes two text boxes:

- One for specifying the location of the interview.
- Another for the city where it took place.

If these fields are left blank, they will automatically assume the default values based on the system settings of the machine running the program.

[37] For testing purposes, fictitious names were created. However, in the final version, the program will be classified as restricted and will include the correct information of the interrogators

On the same row, there is also an informational button that will display “Need Validation” if any of these fields remain unfilled. Once all fields are correctly entered, the button status will update to “Validated”.
Note: The previously mentioned fields are optional and do not prevent the program from proceeding with its analysis. If the data is not entered, the interviewer can manually complete the information later in the Word document.
The PJM icon serves as a shortcut to the PJM website, while the Outlook icon opens the machine’s Outlook application with the existing system account, automatically attaching the interrogation data for easy email forwarding.
In the top left corner, there is a dropdown menu containing the following options:

- Load Video
- Load Audio Recording
- Use Camera
- Exit, which closes the program, functioning the same way as the “X” button in the window.

The “Load Video” option allows users to upload pre-recorded interrogation videos in the following formats: .mp4, .avi, and .mov. Selecting this option opens a file selection window where a compatible file must be chosen.
Similarly, the “Load Audio Recording” option enables the user to upload audio files in .mp3, .wav, and .ogg formats. The process is identical to loading a video.

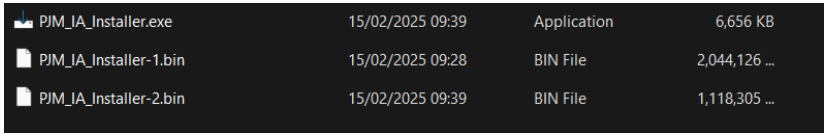
The final option, “Use Camera”, opens a live camera feed and starts recording in real time. To exit this mode, the user must press the “Q” key, which will stop the recording. Alternatively, the user can close the recording window by clicking the “X” button.
Lastly, as previously explained, the bottom left and right sections of the interface contain quantitative emotion analysis graphs. The left-side graph corresponds to FER, which logs an emotion count each time a change occurs. The right-side graph corresponds to SER, which analyses the voice tone in 10-second intervals, recording and classifying the detected emotions.

When the analysis process is complete, and both emotion analysis graphs are displayed, all documents will have been fully generated and can be opened for review.
At the end of the process, all temporary documents created during the analysis are deleted. Simultaneously, a new folder is automatically created inside the “Corpus” directory, named according to the interviewee's ID (Citizen Card number) and the date and time of the interview (e.g., 12345678_YYYY-MM-DD_HH-MM-SS).
Inside this folder, the following fully completed documents will be stored:

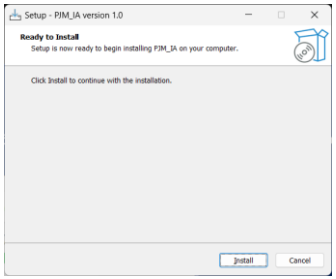
- The three standard reports used by the PJM:
  - “Complaint Report”
  - “Witness Examination Report”
  - “Interrogation Report of Defendant”
- The automatically filled PDF report, including all relevant case details.
- A processed video of the interview, in which emotion classifications are displayed in the bottom left corner at 10-second intervals, visually tracking the subject’s emotional state throughout the session.

Instalation

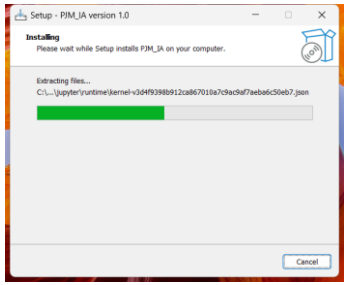
For installation, the program consists of three files, and the user should run the file named PJM_IA_Installer.exe to initiate the installation process.



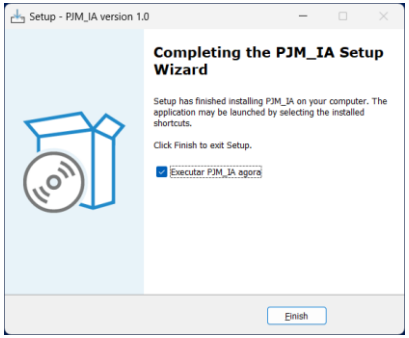
After double-clicking to start the installation, the initial installation window should appear. The user must then click the “Install” button to proceed with the installation.



There is no need to select the installation directory, as the program is automatically installed in the system's program directory by default. The installation process may take some time, so users should allow it to complete fully without interruption.



At the end of the installation, you may choose to run the program immediately or later. To launch it right away, simply select the checkbox next to “Run PJM_IA now” before clicking “Finish”.



When executing the program, two windows will open:

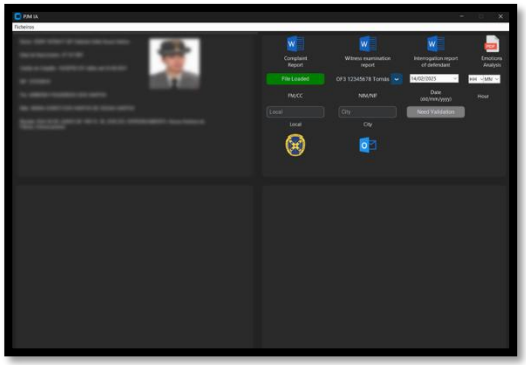
- A command-line window, which must remain open throughout the entire process. If this window is closed, the INTU-AI program will also be terminated.
- The login window, where users must enter their credentials to access the system.

In addition to the installation directory, the program will also create a shortcut on the desktop for quick and easy access.

Usage Mode

The program starts with the authentication process, as mentioned in the Operation Module chapter. In the main menu, the first step is to load the information related to the interviewee. This step is mandatory, and failing to complete it will result in an error.

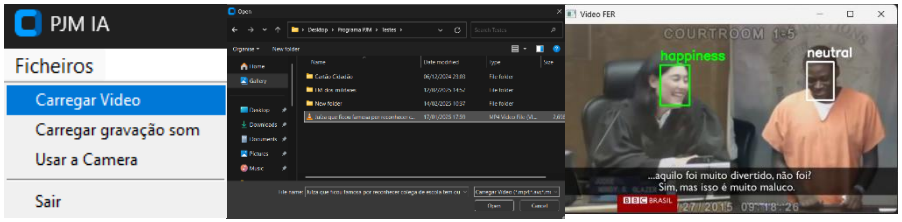
Upon loading the MF or CC, a menu will appear in the upper-left corner, displaying a summary of relevant information to assist the investigator in conducting the interrogation.



As a second step, which is optional, the investigator should enter their details in the combo boxes and text boxes located in the upper-right section of the program. If these fields are left blank, their values will default to “None”, except for the Location and City fields, which will be automatically assigned based on the system's IP address.
Next, the user proceeds to the program's main functionalities. As previously mentioned, the system allows three types of input options: Loading Videos, Loading Audio Files, or Using the Camera for real-time recording.

1. ***Input Video***

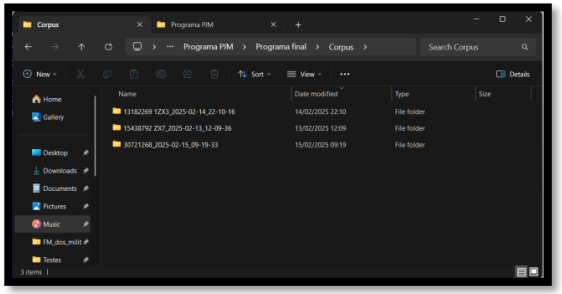
In the upper-left corner, within the dropdown menu, the user should select the option “Load Video.” This will open a file selection window where a video must be chosen in one of the supported formats: .mp4, .avi, or .mov. Once selected, a new window will open displaying the video analysis in progress.
The user should allow the video to play until they reach the desired stopping point for analysis. By default, the program will process the entire video unless manually interrupted.



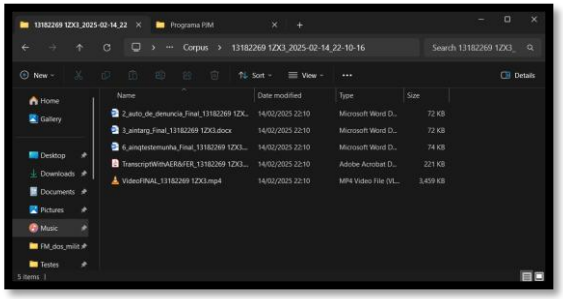
Once the previous process is completed, the FER graph (Facial Emotion Recognition) will appear in the lower-left corner. However, the program will continue processing the remaining data, and the analysis is not yet complete. The process will only be finished once the SER graph (Speech Emotion Recognition) appears in the lower-right corner.
This analysis takes a few minutes, and the user should allow the process to run until the SER graph is displayed, indicating that the full emotional analysis has been successfully completed.

At this point, the documents available in the upper-right corner, namely the three Word documents and the PDF document, are ready and can be opened. Simultaneously, at the end of the previous process, the program automatically creates a folder in the installation directory, specifically within the /Corpus folder. This folder is named according to the interviewee's

identification number (CC) and the date and time of the interview (e.g., 12345678_AAAA-MM-DD_HH-MM-SS).



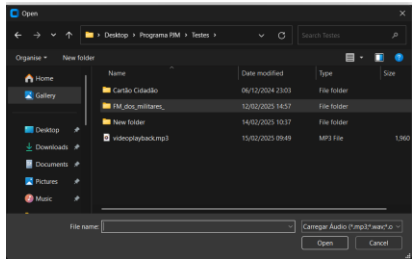
Inside the folder, we can find the three completed Word documents, as well as the PDF document and the processed video. The video includes annotations displaying the detected emotions from FER (Facial Emotion Recognition), SER (Speech Emotion Recognition), and Sentiment Analysis. At this stage, the interrogator should delete any Word files that do not correspond to the interviewee, ensuring that only the relevant documents are retained.



2. ***Input audio***

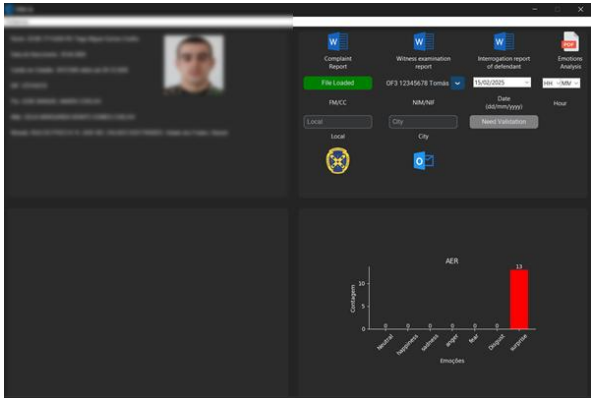
The procedure is identical to the previous one, with the key difference being that in this case, the program does not analyse FER (Facial Emotion Recognition). Instead, it only processes SER (Speech Emotion Recognition).

Apart from this distinction, the rest of the procedure remains exactly the same as before.

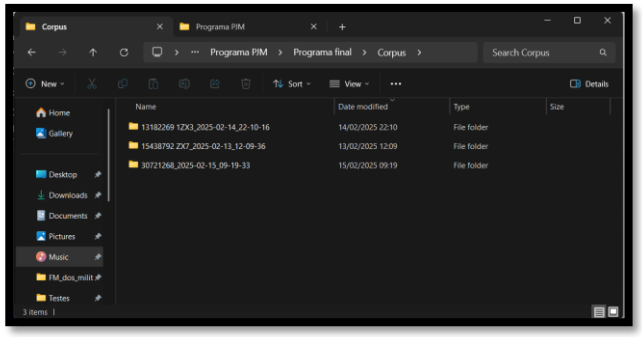


Unlike the previous process, no additional window will open; instead, the program will simply process the audio in the background. This processing time can be considerable, especially for larger audio files.

At the end of the process, a graph related exclusively to SER (Speech Emotion Recognition) will appear in the bottom left corner of the interface. At this point, the documents in the top right corner,namely the three Word documents and the PDF report,will be fully generated and ready to be opened.



Simultaneously, at the end of the previous process, the program automatically creates a folder within the installation directory, specifically inside the /Corpus folder. The folder is named based on the interviewee's identification number (CC) along with the date and time of the interview (e.g., 12345678_AAAA-MM-DD_HH-MM-SS).

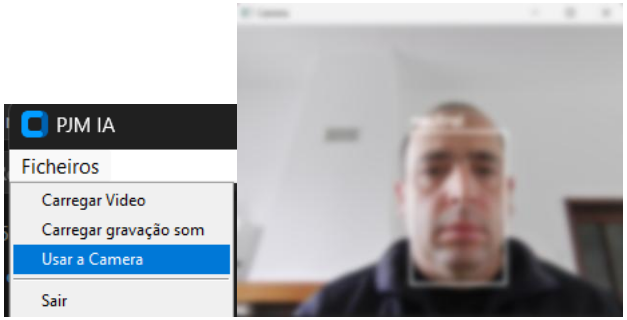


Inside the folder, you will find the three completed Word documents, as well as the PDF report and the audio file used for analysis. Additionally, if a video was processed, it will contain annotations displaying the detected emotions (SER and Sentiment Analysis). At this stage, the investigator must delete any Word files that do not correspond to the interviewee.

3. ***Streaming***

In the streaming or open camera mode, your machine must be equipped with a USB-connected camera as well as a USB-connected microphone. It is essential to comply with the technical requirements to ensure proper functionality.

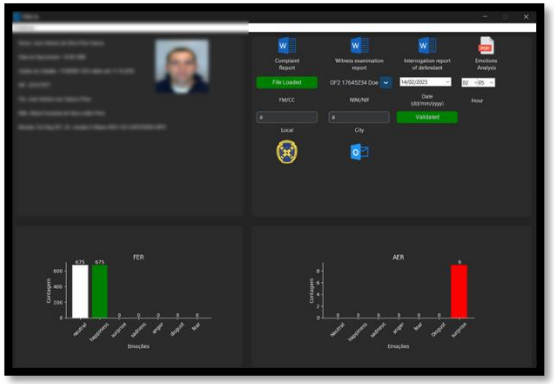
To start the streaming mode, navigate to the cascading menu in the top-left corner and select “Use Camera”. This will open a new window displaying the live camera feed.



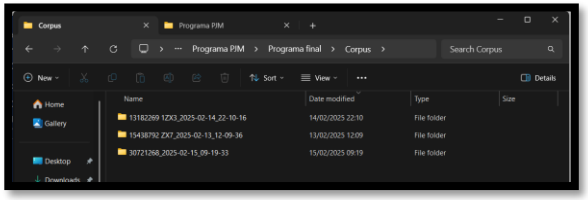
At the end of the recording, you should press the “Q” key. This action will stop and interrupt the recording, after which you should close the camera window manually.

From this point on, the process follows the same steps as loading a pre-recorded video, as explained in section 1. Initially, the FER graph will appear in the bottom-left corner, indicating that the facial emotion recognition process is complete. However, the program will continue processing the remaining data, and the analysis will only be fully completed when the SER graph appears in the bottom-right corner.

This processing phase may take several minutes, and it is important to allow it to run until the SER graph is fully generated.



From this moment, the documents available in the upper-right corner, namely the three Word documents and the PDF report, are ready to be opened.
Simultaneously, at the end of the process, the program creates a folder within the installation directory, specifically inside the /Corpus folder. This folder is named after the interviewee's CC number, along with the date and time of the interview (e.g., 12345678_YYYY-MM-DD_HH-MM-SS).



Inside the folder, you will find the three completed Word documents, as well as the PDF report and the processed video. The video includes annotations displaying the detected emotions (FER, SER, and Sentiment Analysis).
At this stage, the interrogator should delete any Word documents that do not correspond to the interviewee.

9.2. Annex B - User Story – INTU-AI

Title: Investigation Support Program - INTU-IA
User: PJM conducts regular interrogations, and with the rapid advancements in technology, there is a growing need for a tool that automates the analysis of a suspect's emotions, records statements, and streamlines the generation of official reports. This would enhance the efficiency and accuracy of investigations, providing a more structured and data-driven approach to law enforcement.
Scenario: An investigator receives a new case and needs to interrogate a key witness. Currently, the process is manual and time-consuming, requiring video and audio recordings, manual transcription of statements, and the creation of formal reports. The PJM seeks a solution that automates these tasks and provides insights into the witness's emotional state, enhancing the efficiency and accuracy of the investigative process.

Experience Flow with INTU-AI

1. **Login and preparation**: The investigator begin the INTU-AI session by logging in with their credentials. On the main dashboard, they input key details of the interrogation, such as date, time, and location, ensuring that all information is accurately documented.
2. **Video and audio capture:** During the interrogation, the investigator uses the system's camera and microphone to record the interaction with the witness. INTU-AI processes these recordings in real-time, analysing:
    - Facial Expressions (FER)
    - one of voice and emotions (SER)
    - Sentiment of the verbal content (Sentiment Analysis NLP).
3. **Emotion Analysis and Classification:** The system automatically segments the video and audio into 10-second clips and classifies each segment based on the detected emotions. Each clip is then automatically renamed with emotion and sentiment labels, for example: "*clip_0s_10s_anger_sadness.mp3*"
4. **Graphical Visualization of Emotions:** The PJM can monitor real-time statistical graphs displaying the distribution of the witness's emotions throughout the interrogation. This feature enables investigators to identify behavioural patterns and critical moments during the interview, providing valuable insights into emotional fluctuations and inconsistencies that may indicate deception or stress.
5. **Automatic Transcription and Report Generation:** INTU-AI automatically transcribes the audio and generates a summarized report of the interrogation, ready for inclusion in the investigative process. The system ensures that all spoken content is accurately documented, reducing manual transcription efforts and improving efficiency. The officially generated documents include:
    - Complaint Report – A structured report detailing the complaint registered during the investigation.
    - Witness Examination Report – A formal document summarizing the witness's statements and emotional analysis.

- Interrogation Report of the Defendant – A comprehensive report containing the suspect's responses, behavioural patterns, and emotional variations throughout the interrogation.

This automated process not only enhances accuracy but also ensures a standardized and structured documentation workflow for law enforcement agencies.

6. **Synchronization and Playback of Analysed Video**: INTU-AI allows investigators to review the interrogation video while visualizing the detected emotions over time. João can watch the processed video with real-time emotional insights, helping to identify key moments of behavioural changes. The system generates a final video containing interrogation clips with overlaid emotion labels, displayed in three distinct columns:
    - SER (Speech Emotion Recognition) – Captures emotional variations in speech, analysing tone, pitch, and intensity.
    - FER (Facial Emotion Recognition) – Detects facial expressions and micro expressions associated with emotional states.
    - Text Sentiment Analysis – Analyses spoken content for sentiment classification, identifying contradictions between speech and expressed emotions.
7. **Export and Sharing**: With a single click, INTU-AI enables the seamless export and sharing of interrogation data. The system provides multiple options to facilitate collaboration and documentation, ensuring that all generated information is easily accessible and securely stored. The PJM users can:
    - Send Reports via Email – Automatically attach and send the interrogation reports directly from the system.
    - Access Generated Documents – Retrieve official documents in PDF and Word formats, ready for review and inclusion in the investigation file.
8. **Benefits of INTU-AI for Criminal Investigations:**

- **Complete Automation** – Reduces time spent on transcription and emotion analysis, allowing investigators to focus on case-solving.
- **High-Precision Emotion Analysis –** Detects deception cues and inconsistencies in interrogations, enhancing investigative accuracy.
- **Professional Reports** – Automatically generates structured and formatted reports for official use, streamlining documentation.
- **User-Friendly Interface** – Optimized workflow with an intuitive design, making it accessible to all investigators.
- **Seamless Integration** – Audio, video, and text are processed in a single platform, ensuring a comprehensive interrogation analysis.

Conclusion

The INTU-AI program revolutionizes the way the PJM conducts interrogations by providing a powerful AI-driven tool that enhances investigative efficiency, improves the accuracy of emotional analysis, and automatically generates comprehensive reports. By leveraging this technology, PJM can focus its efforts on what truly matters, solving cases more quickly and effectively.

9.3. Annex C – Operational Requirements/Technical Specifications INTU-AI

MINISTÉRIO DA DEFESA NACIONAL
PJM

INTU-IA Software

Technical Specification

November 2024

CHAPTER I - GENERAL TERMS

The INTU-AI (Military Judiciary Police with Artificial Intelligence) program was developed with the aim of creating a technological tool to support criminal investigations, focusing on the automated analysis of interrogations through a multimodal approach based on Artificial Intelligence. This system is designed to assist investigators during interrogations by providing real-time emotional analysis of subjects, automatic transcription and interpretation of verbal content, and the automated generation of official interrogation reports.
The INTU-AI architecture integrates three core modalities:

- Facial Emotion Recognition (FER): Uses computer vision models to detect and classify facial expressions in real-time or from recorded video, identifying basic emotions such as anger, sadness, surprise, fear, among others.
- Speech Emotion Recognition (SER): Analyses acoustic features of the speaker's voice to detect emotional patterns based on intonation, pitch, and vocal rhythm.
- Text-Based Emotion Analysis (NLP): Transcribes interrogation content and applies Natural Language Processing techniques to infer the emotional state of the speaker based on semantic and syntactic structures.

Beyond emotion detection, the system also aims to infer the truthfulness or deception of statements by leveraging emotional patterns identified across the three modalities. This enables a more objective and evidence-based approach to lie detection.

- Designed as an end-to-end solution, the INTU-AI system allows for:
- Input of video, audio, or real-time streaming;
- Automated segmentation of multimodal data;
- Synchronous analysis of all three modalities;
- Generation of annotated reports and videos with emotion insights;
- Exporting results in formats compatible with PJM documentation and procedures.

The INTU-AI platform also features an intuitive user interface, organized into distinct functional sections, enabling both real-time analysis via live camera feed and offline processing of pre-recorded files. It is an autonomous and secure tool, purpose-built for forensic environments, and compliant with principles of confidentiality, data integrity, and the legal admissibility of evidence.

CHAPTER II - OPERATIONAL REQUIREMENT (RO) AND TECHNICAL SPECIFICATION (TS) FOR INTU-AI APLICATION

a. Operational Requirements (OR1) – Support Tool for Interrogations

The system must act as an auxiliary tool to investigators, enhancing the efficiency and effectiveness of interrogation processes.

(1)- Technical Specification (TS1)	(2)- Technical Specification (TS2)	(3)- Technical Specification (TS3)
Usable at Microsoft windows system.	No web soluction.	Must be installed.

b. Operational Requirements (OR2) – Multiformat Input Support

It must process pre-recorded interrogations provided in video format (with or without audio), or audio-only files. This ensures flexibility in handling diverse input sources.

c. Operational Requirements (OR3) – Real-Time Analysis Capability:

The program should enable live interrogation analysis using real-time camera input, supporting dynamic and adaptive investigation processes.

d. Operational Requirements (OR4) – Automatic Report Generation:

One of the key outputs is the automatic filling of official PJM interrogation reports. This feature streamlines documentation tasks and minimizes human error.

(1)- Technical Specification (TS1)	(2)- Technical Specification (TS2)	(3)- Technical Specification (TS3)
Complaint Report.	Witness Examination Report	Interrogation Report of Defendant

e. Operational Requirements (OR4) – Visual Identification:

To ensure data integrity and correct association between data and individuals, a facial identification system is required to confirm the identity of the interrogator versus the suspect.

f. Operational Requirements (OR5) – Document Type Recognition:

The system must detect and differentiate between various types of documents (e.g., National ID Card or Military Form) to extract relevant information accordingly.

(1) Technical Specification (TS1)

The military form it's a number with mandatorily eight-digits.

Note: Normally numbers like 00012345 are represented by 12345

(2) Technical Specification (TS2)

The National ID Card it's a number with mandatorily eight-digits followed with a four -digit code (more information about it can be consulted in national reference).

g. Operational Requirements (OR6) – Robust Data Handling:

It must handle large volumes of data and cases of missing data without crashing or producing invalid results.

h. Operational Requirements (OR7) – Live Emotion Recognition:

The program should support real-time emotion detection during live interrogations, providing an extra layer of behavioural analysis.

i. Operational Requirements (OR8) – Video Upload for Asynchronous Analysis:

Users must be able to upload recorded videos for offline analysis.

j. Operational Requirements (OR9) – Window Management for Real-Time Streaming:

The camera and main windows must be decoupled to allow viewing of suspect data in one window while monitoring the real-time video feed in another.

k. Operational Requirements (OR10) – Audio-Only Input Support:

If only audio is provided, the system should output detailed emotion and sentiment reports.

l. Operational Requirements (OR11) – End-to-End Automation:

From data acquisition (video, audio, or live stream) to processing (FER, SER, emotion analysis), data synchronization, document generation, storage, and sharing, the system must operate as a fully automated pipeline.

m. Operational Requirements (OR12) – Secure Access:

The system must implement user authentication via a login screen, ensuring only authorized personnel access the data.

(1)- Technical Specification (TS1)	**(2)- Technical Specification (TS2)**
Must have a system to prevent error, or when the user doesn't introduce the right password or username must have some kind of information, granting the user the information of that fact.	The user window when closed must kill the program, preventing the user to enter without authentication.

n. Operational Requirements (OR13) – Main Menu Functionalities:

The interface should allow video/audio uploads, report access, and manual entry of relevant information.

o. Operational Requirements (OR14) – Streaming Exit Option:

While using the real-time camera function, the user must be able to stop the process using an exit button.

p. Operational Requirements (OR15) – Structured Interface Layout:

The main menu is divided into four quadrants, each supporting specific functionalities.

(1)- Technical Specification (TS1)	(2)- Technical Specification (TS2)	(3)- Technical Specification (TS3)
Two quadrants with SER and FER resume for the interview.	One quadrant with a textbox/combo box combination, to ensure some information. And a button system to navigate to the reports.	One quadrant must have the information at OR4, manly photography, name, birthdate, father's and mother's name, address.

q. Operational Requirements (OR16) – Graphical User Interface:

Must be built in a windows approach, must have de geometry of 1200x800.

(1)- Technical Specification (TS1)	(2)- Technical Specification (TS2)	(3)- Technical Specification (TS3)
Must be built with dark them colours for layout.	The graphic of SER and FER must have distinct colours, and easy identified colours.	The coulors of the SER, FER and Emotion analysis must have the same logic.

r. Operational Requirements (OR17) – Document Processing

Must have a direct way to the reports in word, and for the report analysis in PDF.

(1) Technical Specification (TS1)
Must respect the reference legal documentation (omitted purposely)

(2) Technical Specification (TS2)
When creating reports, you must eliminate all information relating to the person being questioned and the interrogator.

(3) Technical Specification (TS3)
The interrogation information must be stored in the PJM's own database.

(4) Technical Specification (TS4)
The analysis report must contain on the first page a summary of the entire interview.

s. Operational Requirements (OR17) – Distribution

Delivered as executable file (.exe) and a user manual in Portuguese and English.

t. Operational Requirements (OR18) – Live Emotion Overlay:

Display of detected emotions over the video in real time.

9.4. Annex D – Business Understanding

This work is grounded in the operational context of a highly specialized criminal investigation authority the PJM, is tasked with supporting judicial authorities in criminal investigations and conducting preventive or investigative actions within its jurisdiction, or in cooperation with other judicial authorities.

The present study focuses specifically on interrogations, which currently remain a fully human-driven process. Typically, an interrogation involves a verbal exchange between two[38] main participants: the investigator (inspector) and the interrogated individual.

This procedure encompasses a series of tasks, including preparing the interrogation, conducting the session, recording[39] the suspect's statements, analysing the interaction, and ultimately drafting an official report. The process is time-consuming, and due to its judicial nature, it requires meticulous attention and care.

Currently, the workflow can be described as follows: the interrogation is conducted between the interrogator and the suspect, often with audio recording. In some cases, video recordings are also made when technically feasible. After the session, the investigator is responsible for drafting a comprehensive report, which may include seeking expert assistance in facial expression analysis to interpret the suspect's behaviour and demeanour during the session.

The report itself serves as a formal summary of the interrogation, highlighting key points and incorporating a full transcription of the conversation. It also includes relevant identification details for both the interrogator and the interrogated individual. Given its official nature and importance in judicial proceedings, the report demands full concentration and precision, making it a labour-intensive and sensitive task.

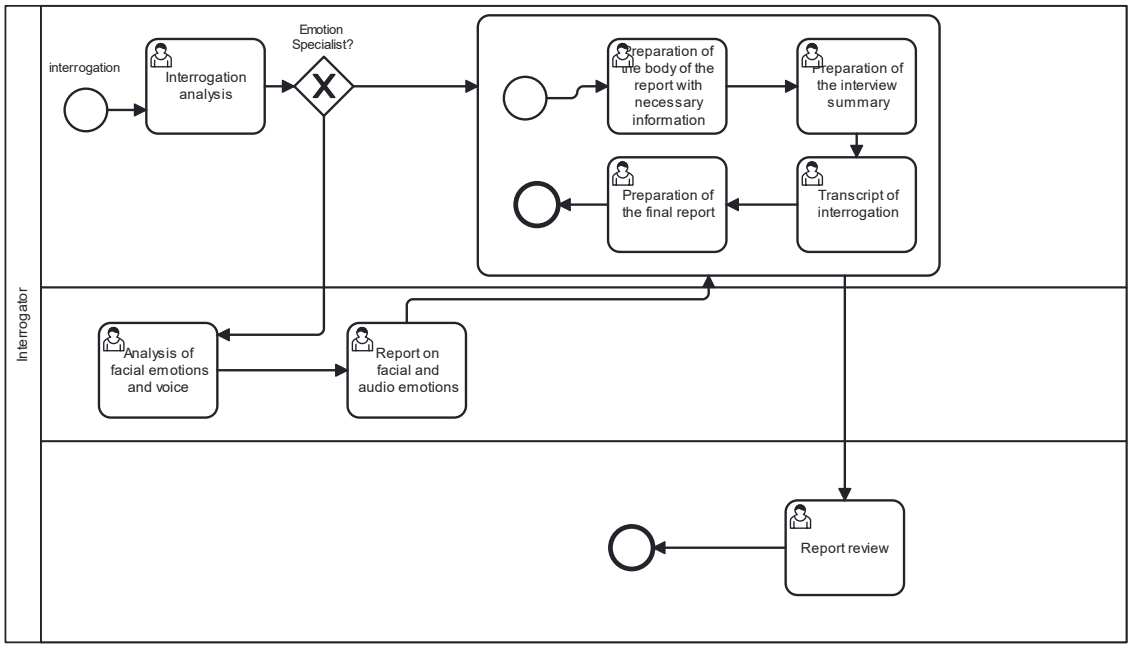


Figure 29 - BPMN process - As-Is Process of Interrogation

With the rapid advancement of technology, this law enforcement agency must keep pace with emerging tools that serve as process facilitators. This need is further intensified by the

[38] It could be more than two interrogators
[39] Audio or Audio and Video

declining number of investigators relative to the volume of ongoing work, often leading to human error due to time pressure and workload.
Therefore, the ultimate goal of this project is to empower the PJM with an AI-driven tool capable of assisting investigators during interrogations. Conceptually, the objective is to minimize human intervention in much of the interrogation workflow, enabling greater efficiency, consistency, and accuracy in both analysis and documentation.

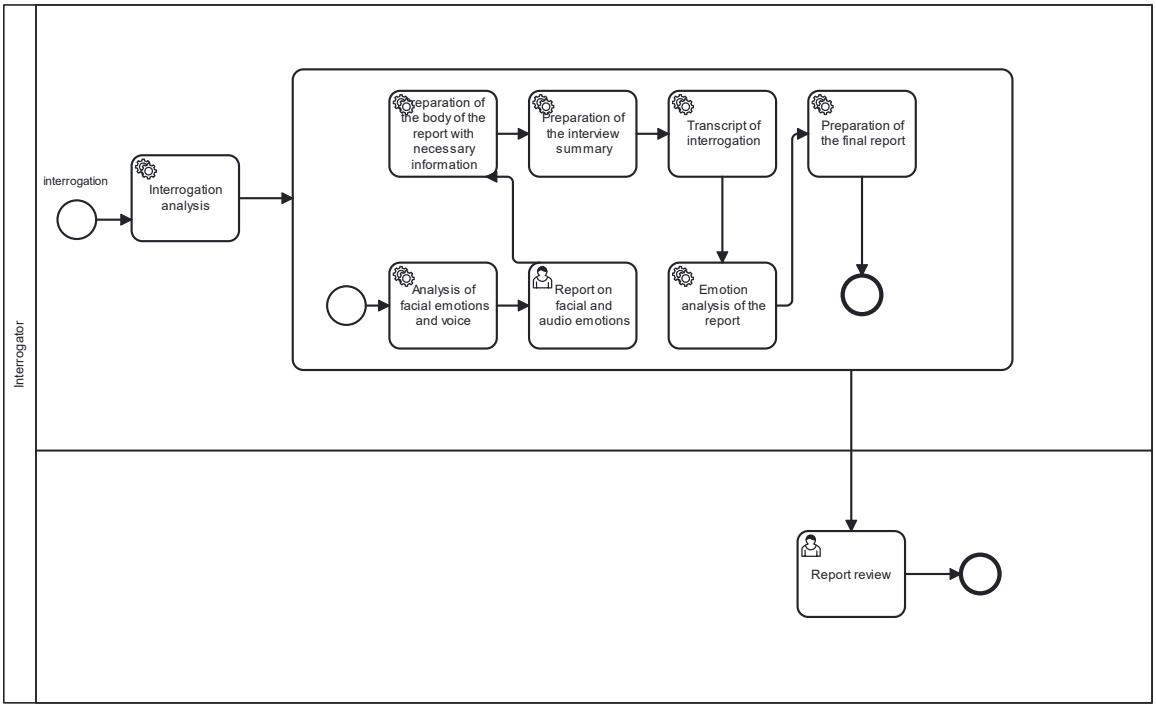


Figure 30 - BPMN process - To-Be Process of Interrogation

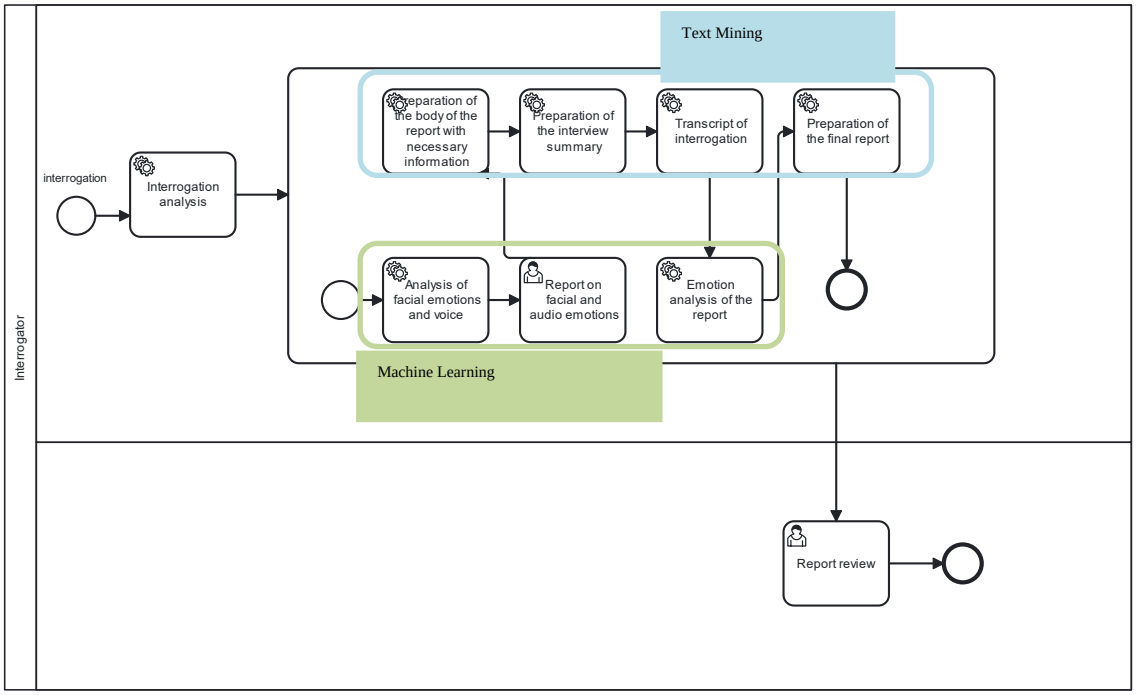
9.4.1. Determine business objectives

The tool being developed is not intended to replace the role of the investigator in their work, but rather to support investigations and streamline the overall process. In this context, and in parallel with the functionalities presented so far, the program must also be capable of performing sentiment analysis on the statements made by the interviewee.
The solution can therefore be divided into two main areas of operation:

- Machine Learning – Focused on the design and training of emotion classification models, whether through facial expressions (FER), voice signals (SER), or text (Text-based Emotion Analysis).
- Text Mining with NLP – Dedicated to processing the textual transcriptions of the interrogations using Natural Language Processing techniques such as:
    - Information Extraction
    - Summarization
    - Categorization

Together, these components aim to provide a comprehensive, intelligent, and adaptive tool that enhances the investigator's capabilities, reduces manual workload, and improves the efficiency and accuracy of the interrogation analysis process.

Table 17 - BPMN process with application areas



1st Objective of the Organization:

Automate the administrative process of completing reports resulting from interrogations, including the population of data related to both the interviewers and interviewees, as well as the transcription and summarization of the interrogation content.

2nd Objective of the Organization:

Perform emotional analysis of interrogations, either in real time or retrospectively, by identifying the emotional states of the interviewee throughout the session.

3rd Objective of the Organization:

Generate a summary report of the emotional dynamics during the interrogation, along with an annotated video indicating the emotions detected over time.

4th Objective of the Organization:

Associate the detected emotions with specific segments of the interrogation timeline in order to support investigative analysis and decision-making.

9.4.2. Assess the scenario

In our Information Extraction challenge, one of the technical specifications of the project involves the identification and transcription of speech associated with specific individuals. While it is desirable (though not mandatory), the goal is to segment the transcribed text according to the speaker, distinguishing between interrogator and suspect.

In the context of Information Extraction, it is a requirement that the system should only accept valid Portuguese national identification documents: MF or CC. In cases where the interrogated person does not possess either document, this scenario applies exclusively to the PJ, as the PJM deals solely with military personnel, the program must allow for manual editing of the generated report to insert the missing identification data. It was agreed that features such as automatic recognition of alternative documents (e.g., passport) may be implemented in a future version.
The following information must be extracted from these documents (both for on-screen visualization and automatic report generation):

Table 18 - Information to be extracted from identification data

Information to extract from identification data		
Full name	NIM[40]	Rank[41]
Date of birth	CC and expiration date	
NIF	Father's and Mother's full name	
Full address (street, postal code, city)	Photograph of the interrogated person	

9.4.3. Objectives of TM and ML in Model Development

9.4.3.1. Text Mining Objectives

Aligned with the first and second organizational goals, Text Mining plays a crucial role in processing, structuring, and understanding textual data derived from interrogations:

- Automatic Transcription: Convert spoken language from audio/video recordings into accurate written text using advanced speech-to-text models (e.g., Whisper).
- Information Extraction: Automatically extract key information such as full names, ranks, IDs, addresses, and other relevant personal data from transcriptions and identification documents.
- Text Summarization: Provide concise summaries of the full interrogation to reduce cognitive load on investigators and streamline report writing.
- Sentiment and Emotion Classification: Apply NLP-based models (e.g., RoBERTa, BERT) to detect emotional tone from the transcribed statements.

9.4.3.2. Machine Learning Objectives

Fulfilling the second, third, and fourth organizational goals requires the deployment of emotion recognition capabilities across multiple modalities:

- FER: Utilize deep learning models to detect facial expressions in video frames and classify them according to basic emotions (anger, fear, joy, etc.).
- SER: Analyse vocal features using models such as CNNs or transformers trained to detect emotional states in spoken language.

[40] For Military personnel
[41] For Military personnel

- Text-based Emotion Recognition: Apply transformer-based NLP models and traditional models as, NB, Logistic Regression and others to infer emotional content from the suspect’s spoken words, completing the triad of emotion analysis.
- Multimodal Fusion: Integrate FER, SER, and text-based emotion predictions through late fusion techniques, combining their outputs into a unified and more robust emotion classification model.