Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Judgments Of Emotional Reactions By Facial Emotion Recognition System: A Comparison

Rishi Malpani

Supervisor: Khurshid Ahmad

February 12, 2024

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Masters in Science Computer Science

Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

I agree that this thesis will not be publicly available, but will be available to TCD staff and students in the University's open access institutional repository on the Trinity domain only, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: _____ Date: _____

Abstract

The research begins by asserting that a change in the movement of the muscle group responsible for executing a facial action unit permits one to ascertain the person's emotional state. The connection between muscle movements and emotions is what makes it possible to build up a recognition system. Emotion has physical correlates that are independent of race, culture, and age. We looked at how two systems recognize emotions by watching videos of people showing different emotions, both real and imagined. We also looked at videos of people who tend to be good at controlling their emotions, like politicians and leaders. Furthermore, we demonstrate that there is a difference in the emotion judgments by two major emotion recognition systems, Emotient and Affectiva, in both posed, spontaneous, and semi-spontaneous emotions, which we have traced down to the level of action unit in that emotion measurements can be correlated to the difference in the measurements of action units. This can be attributed to the difference in algorithms of the two systems. Furthermore, PosedDataset Ravdess and Spontaneous Dataset (AM-FED) and semi-spontaneous datasets from our collection (Politicians and Governors) baselines were used. Can emotion recognition systems with different architectures, training, and testing methods have consistent emotion detection results? The reliability and variation of the emotion recognition results on the spontaneous, semi-spontaneous, and posed database were examined using statistical techniques, such as the Spearman correlation coefficient, Kruskal-Wallis tests, Chi-square tests, and Pearson tests.

Acknowledgements

I want to start by giving a big thanks to my supervisor, Professor Khurshid Ahmad. He's been a fantastic guide and a real source of motivation throughout my M.Sc. degree. I can't even begin to express how much time and energy he's put into helping me with my dissertations. Working with you, Professor, has been a real honour, and I truly hope I've met your expectations and lived up to the trust and confidence you've had in me. I'd also like to express my gratitude to Professor Carl Vogel. Thank you, Professor, for your valuable advice and insights, especially when it comes to tackling the complex world of statistics. Your guidance has been a beacon in helping me find my way through this challenging maze. I want to extend my heartfelt thanks to my teammates in the Research Lab Sarathkumar, Deepayan Datta , Subishi Chemmarathil Your willingness to share your data sets and offer valuable insights on various topics whenever I needed them meant a lot. I would like to thank a Ph.D. student in my Research Lab Darren Scott for making the time spent at the lab so memorable. I would like to thank my mother Aparna Malpani my father Anil Malpani and my sister Riya Malpani for their unwavering support and encouragement throughout this journey, especially during the toughest times. Your blessings and constant encouragement have been instrumental in making me a successful researcher. I want to express my heartfelt gratitude to my parents for their countless and often unspoken sacrifices, ensuring that I could pursue my dream of Trinity College Dublin. Their love and blessings have been a constant source of strength. I would also like to extend a special thank you to Trinity College Dublin for providing me with the opportunity to pursue my chosen field. Your support has been instrumental in shaping my path, and I am truly thankful. Thanks, Everyone!

Contents

List of Figures

List of Tables

1 Introduction

1.1 Emotion Detection in Facial Expression

The field of artificial intelligence holds a significant position in the field of emotion detection, where individuals are capable of detecting an individual's emotions through their facial expressions, gestures, voice, and speech content. For machines, emotion recognition systems are used to detect emotions expressed through non-verbal communication, and textual sentiment analysis techniques can be used to analyse the emotions expressed through verbal communication. Common emotion recognition systems include facial emotion recognition (FER) systems. AFFDEX, a system developed by Affectiva, and FACET, a system developed by Emotient, are two commercial FER systems. The two systems are based on the Facial Action Coding System (FACS), developed by Ekman and Friesen [2]. Through face recognition and automatic facial expression recognition technology, both systems find action units on the human face and make predictions about emotions based on them. Both systems can recognize emotions well, according to previous studies. Emotient has been acquired by Apple, and both systems are available through the IMotions Software [3]. Both systems employ a similar approach to emotion expression, based on the work of Paul Ekman. Both systems employ machine learning technology.

Applications Of Facial Expressions

Facial emotion detection has a wide range of applications.

1. Driver Monitoring: Facial emotion detection can be used to monitor the drivers' emotional state in real time. Cameras or sensors can look at facial expressions and see if the driver is tired, distracted, stressed, or drowsy. It is possible to use this information to alert the driver or trigger automated safety systems to enhance overall safety on the road.

2. Many recruiters would appreciate the technology that can detect a candidate's emotional state and whether they are being honest or concealing their emotions.

1.2 Problem statement

Vogel and Ahmad (2023) [4] have noted that there are differences in the output of emotion measurement systems given the same input. The authors compared emotion-related outputs, which are called happy, sad, angry, surprised, contemptuous, or fearful, and found that each emotion-related output was statistically different. However, each emotion is computed by combining the weighted average of the activation of muscles (action units). I have investigated the differences in the activation of action units.

I have gathered the videos of speeches delivered by politicians and governors of state banks. My data contains individuals of various ages, races, and genders. To combine a baseline with semi-spontaneous videos, I have used two posed databases, AMFED+ [5] and Ravdess [6].

1.3 FACS Action Units and the association with emotion

To facilitate the research and application of human facial expressions, the Facial Action Coding System (FACS) was proposed. The first attempt to encode facial expressions was made in 1969 by the Swedish anatomist Hjortsjö, and he encoded 23 facial expressions. Subsequently, Ekman and Friesen established FACS based on Hjortsjö's work in 1976. Eventually, Ekman updated FACS further in 2002. The three types of action units now employed in FACS are the primary action units, head motion action units, and eye position action units. The facial muscles associated with each action unit may be single or multiple. These action units can encode each individual facial motion. Table 1.1 shows some of the action units and the possible emotions associated with them([7] , [8] , [9]).

Action Unit	1	2	3	4	5	7	9	10	11	12	14	15	20	23	24	25
Emotions																
Author [S.Du] [7]																
Happy										✓						✓
Sad				✓								✓				
Fear	✓			✓									✓			✓
Angry				✓		✓									✓	
Surprise	✓	✓														✓
Disgusted							✓	✓								
Author [P.Lucey] [8]																
Happy										✓						
Sad	✓			✓					✓			✓				
Fear	✓	✓		✓												
Angry														✓	✓	
Surprised	✓	✓			✓											
Disgusted							✓	✓								

Table 1.1: Emotion Description in terms of facial action units

We see that when author s.du [7] wrote about their databases to train emotient, they found out that very few times action units could be associated directly with their sentiments, except happy, and the associations between emotions and action units are different for different systems. The author p.lucey [8] wrote about their data to train affectiva, they found out most of the emotions associated with different action units at the same time except emotion, happy . Table 1.1 indicates that the emotion description in terms of facial action units with sentiments is not exactly one-to-one. Additionally, the table provides a quick reference to see which action units are associated with each emotion, according to the two mentioned authors. It appears that there is some variation between the author's findings, as certain action units are associated with specific emotions in one study and not in the other. This could be attributable to disparities in methodologies, datasets, or other factors that may have an impact on the author's conclusions. There are a few action units that participate in multiple emotions, and there are a few that are unique to each other.

Action Units & Description of Muscle Movements

Action Units	Description of muscle movement	Description of muscle movement
1	Frontalis, pars medialis	Inner corner of eyebrow raised
2	Frontalis, pars lateralis	outer corner of eyebrow raised
4	Depressor Supercilli, Currugator	Eyebrows drawn medially and down
5	Levator palpebrae superioris	Eyes Widened
6	Orbicularis oculi, pars orbitalis	Cheeks raised, eyes narrowed
7	Orbicularis oculi, pars palpebralis	Lower eyelid raised
9	Levator labii superioris alaquae nasi	Upper lip raised and inverted; superior part of the nasolabial furrow deepened; nostril dilated by the medial slip of the muscle
10	Levator Labii Superioris	Upper lip raised;nasolabial furrow deepened producing square-like furrows around nostrils
11	Levator anguli oris	Lower to medial part of the nasolabial furrow deepened
12	Zygomatic Major	Lip corners pulled up and laterally
13	Zygomatic Minor	Angle of the mouth elevated; only muscle in the deep layer of muscles that opens the lips
14	Buccinator	lip corners tightened.Cheeks compressed against teeth
15	Depressor anguli oris (Triangularis)	Corner of the mouth pulled downward and inward
16	Depressor labii inferioris	Lower lip pulled down and laterally
17	Mentalis	Skin of chin elevated
18	Incisivii labii superioris and Incisivii labii inferioris	Lips pursed
20	Risorius	Lip corners pulled laterally
22	Orbicularis oris	Lips everted(funneled)
23	Orbicularis oris	lip tightened
24	Orbicularis oris	Lips pressed together
25	Depressor labii inferioris , or relaxation of mentalis,or orbicularis oris	Lips parted
26	Masseter relaxed temporal and internal pterygoid	Jaw dropped
27	Pterygoids and digastric	Mouth stretched open

Table 1.2: Facial Action Units (AUs)

Table 1.2 describes each action units and descriptions of muscle movements

Muscle Groups and their Subgroups

Main Groups	Subgroups
Muscles of the Mouth (Buccolabial Group)	• Muscles for Elevating and Everting the Upper Lip • Muscles for Depressing and Everting the Lower Lip • Muscles for Closing the Lips • Muscles for Compressing the Cheek
Muscles of the Nose (Nasal Group)	• Levator Labii Superioris Alaeque Nasi - a slender, strap-like muscle found on both sides of the nose • Mentalis - a short conical muscle located in the chin area
Muscles of the Eyelid (Orbital Group)	• Orbicularis Oculi - a sphincter-like muscle that encircles the orbit and the periorbital area • Corrugator Supercilii - a slender muscle found deep to the medial end of the eyebrows
Muscles of the Cranium and Neck (Epicranial Group)	• Occipitofrontalis - a wide muscle that overlies the superior surface of the scalp
Muscles of the External Ear (Auricular Group)	• Auricular Muscles - thin, fan-shaped muscles that connect the auricle to the scalp

Table 1.3: Muscle Groups and Their Subgroups

1.4 Understanding Mental States: Physical Correlations and Microexpressions

Despite the inability to immediately perceive what another person is thinking or feeling, individuals are nevertheless capable of evaluating them (other members' behavior) and discerning what is happening inside their minds. Mind reading, which is also known as indirect inference, deduction, and guess work. We rely on diverse data about how people behave, how they look, how they move, and other factors to draw inferences about their motivations, aspirations, and personalities [10]. Meanwhile, R. El Kaliouby and P. Robinson, in their article, emphasize the importance of expanding the scope of mental state recognition beyond basic emotions and highlight the potential of using physical correlates, such as facial expressions and head gestures, to infer complex mental states in real-time [11].

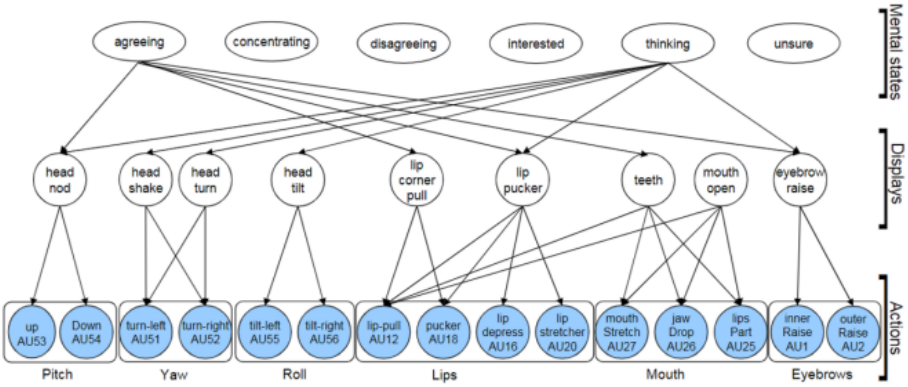

Figure 1.1: Physical correlation of mental states
[11]

A video stream is abstracted spatially into head pitch , yaw and roll actions, and lips , mouth and eyebrow actions.The action are in turn abstracted into displays and mental states.The displays present in a model of a mental state are determined by a feature selection mechanism [11].Figure 1.1 summarizes the spatial abstractions currently supported by the model: head rotation along each of the three rotation axes (pitch, yaw, and roll) and facial components (lips, mouth and eyebrows) [12]

1.5 Structure of the dissertation and my contributions

We will look at how Emotient and Affectiva work at the muscular level of facial expression.Emotient and Affectiva are given a video and then the system makes a frame-by-frame value for each of the six or seven emotions. (Anger, Contempt, Disgust, Joy, Fear, Sadness, Surprise) together with the probability of the activation of each of the 27 or action units. The performance comparison usually is based on the emotional probabilities produced by the two systems. This investigation will be conducted at the level of the activation unit. Note that during the training of the emotion recognition systems, labeled videos(governors/politicans) are used for training. The label for each frame is regressed against the number of action units that were activated over the length of the videos. Using techniques of multi-variate analysis, like independent component analysis and factor analysis, we can figure out what emotions are most active. The regression coefficient is stored and then used to calculate the probabilities of each of the six or seven emotions on the face by adding up all the activations weighted by the regression coefficient [13]. Machine learning techniques are used for regression analysis. Our comparison is based on the 98 videos of facial expressions of political, financial, and sports leaders for each of the leaders we have between 4–5 videos. These are semi-spontaneous videos given at political rallies, special meetings, and sports events. Furthermore, we have used two gold-standard databases comprising 1059 videos. The posed videos are referred to as RAVDESS [6] and spontaneous videos are referred to as AM-FED+ [14]. The results of our investigation reveal statistically significant variations in the outputs of these systems at the activation levels of the action units. We believe that the judgments of emotional reactions by facial emotion recognition may be due to the differences in the training data sets and in the manner in which the face is recognized by the two systems. The videos may be processed to find the activation of the action units [7]. Thus, we conclude that the differences in the

training dataset for the judgment on emotional reactions by facial emotion recognition are due to those two factors. This can help in finding out which action units are activated by these videos [7]. According to different suggestions, there could be about 16 flat muscles for 5 controlling facial deformation. The idea is that different parts of the face are not separate and therefore we cannot exclude activity in one area will have no effect on activity in other regions. Automatic emotion recognition aims at learning the relationship between muscle movements and a person's emotional condition. In Chapter 2, we examine emotion recognition systems employing the method of analysis proposed in this dissertation. In Chapter 3 we discuss how to select the data and the various statistical methods employed.

1.6 Conclusion

The path of progress in the field of emotion recognition spans the entire spectrum, starting with the initial concept of Action Units (AUs) by early researchers like Hjortsjö, and culminating in the current day Facial Action Coding System (FACS) developed by Ekman and others. A significant advancement towards advanced techniques involves the incorporation of machine learning based on labelled data in the implementation of Automatic Facial Coding Systems. Nevertheless, the fundamental problem of reaching complete accuracy in AU detection remains, as accuracy rates vary among the systems. The issues are exacerbated further when it comes to evaluating performance by author Barrett et al [15] especially when it comes to posed expressions, and they present a complicated reality that is not completely simplified when it comes to achieving reliable recognition.

2 Literature Review

2.1 Automatic Emotion Recognition Using Facial Expression

Facial expressions were previously detected through manual FACS coding by humans. However, this method was arduous and slowed down [16]. A normal level of proficiency in FACS required well over one hundred hours of training, and it took an hour to carefully evaluate each minute in the videotape. Following the article titled "Classifying facial actions," [16], there have been numerous advancements in the field of Automatic Facial Emotion Recognition Systems. The foundational insights from this research have been used by systems like FACET EMOTIENT, AFFDEX AFFECTIVA, to harness the Facial Action Coding System (FACS) as a basis for identifying emotions.

Pipline	Emotient	Affectiva
Face Recognition	Viola-Jones algorithm	Viola-Jones algorithm
Feature Detection	10 features	34 features
Classification	SVM	SVM
Training Database	Lab Sourced	Crowd Sourced
Training Dataset	10000 Frames	8000 Frames

Table 2.1: Differences between two systems

FACET EMOTIENT was originally developed by Emotient and later incorporated into the iMotions software suite, This technology originated as a result of a spin-off from the University of California San Diego [15]. However, in 2017, Apple Inc. acquired Emotient. As a result, Facet is no longer available for commercial purchase, but iMotions continues to provide support for existing licenses. Emotient(FACET) tracks facial landmarks and assigns emotion scores on a scale from -1 to 1, representing the intensity of expressed emotions. Emotient facial expression analysis engine also known as FACET (formerly the computer expression recognition toolbox (CERT), is the real technology behind the engine used for processing the videos. It automatically codes the 20 different facial action units and 7 different prototypical facial expressions into emotions. [1]. It also estimates the locations of 10 facial features as well as the 3-D orientation (yaw, pitch, roll) of the head.

AFFECTIVA(AFFDEX) Affectiva a spin-off company resulting from research activities at the MIT Media Lab in 2009 [17], this technology is currently distributed by Smart Eye (through their API and SDK) as well as by IMotions . In the paper [18] they presented the AFFDEX software development kit (SDK). The SDK provides an easy interface for processing multiple

faces within a video or live stream in real-time.AFFDEX is one of the pioneers in the field of Emotional AI. AFFDEX performs emotional analysis thereby scoring emotions (here eight emotions are considered) in a range between 0 and 100 with a valence score ranging between -100 and 100. The product is based on the Viola-Jones algorithm for facial recognition[19]. Here, valence means the intensity of the emotions, i.e.how positive or negative the emotions are, with 0 being the neutral value.

2.2 Review of the comparative studies

Agreement and disagreement between major emotion recognition systems

According to Vogel and Ahmad [4], the evaluation of emotion recognition systems has been a subject of debate and complexity for ages. Early pioneers in this field had big ideas about systems that could detect emotions like anger, happiness, disgust, fear, and sadness from a person's face or voice. The FER systems rely on 'labelled' databases, where videos and audio clips showcase individuals expressing specific emotional states. The machine learning algorithms then compare the pixels in the distribution or waveform to these labels, claiming to learn how to recognize human emotions. They are used extensively, from aiding autistic spectrum communications to surveillance applications, but their training data often comes from idealized conditions - people facing cameras and using noise-canceling microphones. The problem comes when these systems are used with real-life data. Things like lighting, how someone looks, and who they are can affect how well they work. The authors have produced a dataset that includes videos, and their analysis covers soundtracks from 64 politicians and 7 government spokespersons, encompassing diverse demographics such as age, gender, and race. The dataset contains 16.66 hours of data. The authors evaluate two FERs systems, namely Emotient and Affectiva, by analysing their emotional assessments on a frame-by-frame basis. The author's analysis covers head and facial muscle movements and vocal tract muscle movements. They observed marked differences in emotions recognized, with more pronounced differences between women and men and between South and East Asians and White Europeans. These insights provide insight into the levels of agreement and disagreement, both in high-level emotion labels and lower-level features like Euler angles of head movement. Furthermore, they propose that inter-system disagreement could be used as a useful measure for identifying data characteristics that influence disagreement. But it's important to know that there are limits. This work is still ongoing, and they haven't looked at how much agreement there is about low-level features in FER, facial landmark tracking in FERs, and the full interactions between these measurements and other categories. Even though their dataset is growing, it's still small. Additionally, they have not introduced independent human emotion assessments, since our focus is on evaluating the potential for using these systems on data outside their training sets through their agreements and disagreements. To summarize, this paper contributes to the ongoing discussion regarding fairness and bias in machine learning systems by utilizing FERs as exemplary instances. By figuring out what makes emotion recognition systems work, they want to make it easier to use them and make sure they work well in different situations.

2.3 Conclusion

Vogel and Ahmad study explores the details involved in facial emotion recognition systems (FERs). Through testing of two pieces of software, Emotient and Affectiva, they discovered some noticeable differences in facial emotion recognition performance. This investigation explains that the actual implementation of these devices encounters numerous obstacles, which is why they should be assessed by experts with meticulous consideration and considering any unforeseen circumstances. The research suggests that intra-system disagreement could help understand what determines divergence in data features. The hypothesis that speaker-independent claims are possible is challenged by demonstrating notable differences in emotion recognition between clearly and unintelligible speech. Age and ethnicity are also variables that differ, highlighting how the training data, algorithms, and vocal characteristics impact the outcomes. Moreover, the significant disparities in the acknowledgment of certain emotion categories indicate the intricate nature of emotion recognition systems.

3 METHODS

3.1 Methods for data selection

This study looks at how people recognize emotions on their faces. The experimental process includes video collection, video pre-processing, video analysis, data analysis, and hypothesis testing. In this study, we intend to synthesize the sources of emotion using software systems and to analyze the outputs of these systems using statistical inference. Therefore, a pipeline is created to implement the proposed system.

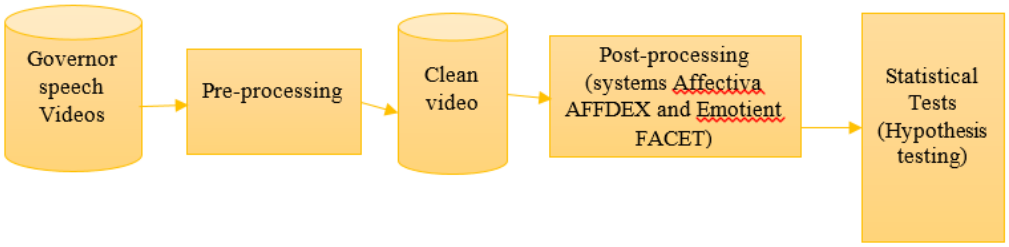

Figure 3.1: System Pipline

3.2 Data Pre-processing

The process starts with the video file (MP4 format). During pre-processing, video files are trimmed and cropped to improve automatic emotion recognition accuracy by reducing the possibility of errors caused by multiple faces or voices, and objects obstructing faces. After selecting the videos from YouTube, these mp4 videos underwent a process of trimming and cropping using an online tool https://online-video-cutter.com/crop-video and Windows Video Editor. The video is zoomed to ensure that there is only one face in the frame. The video files are loaded into the IMotion software package to classify emotions in both AFFECTIVA (AFFDEX) and EMOTIENT (FACET). The output from both these systems is downloaded as a CSV file for further processing.

3.3 Available Data

Spontaneous and Posed Facial Expressions (or "Gold Standard") Dataset

The performance of the facial emotion recognition systems has been evaluated using two publicly available datasets (RAVDESS, AM-FED+). The existing literature for both the gold standard database((RAVDESS, AM-FED+) frequently evaluates emotion recognition systems on posed and spontaneous expression databases, resulting in the set of speeches by politicians and governors representing a semi-spontaneous collection.RAVDESS represents a posed collection, while AM-FED+ represents a spontaneous collection. These datasets were subsequently processed by EMOTIENT and AFFECTIVA.

3.3.1 Collection of Labelled data-sets and Motivation

AM-FED+

An Extended Dataset of Naturalistic Facial Expressions [5], collected in Everyday Setting which contains 1,044 videos of which 545 videos (263,705 frames or 21,859 seconds) have been comprehensively manually coded for facial action units. These videos act as a challenging benchmark for automated facial coding systems. All the videos contain gender labels and a large subset (77 percent) contain age and country information. The data collection protocol was approved by the MIT Committee. The videos from the webcams were streamed in real-time at 14 frames/second and a resolution of 320 X 240.

For the dissertation, 448 videos were selected from 545. The AM-FED+ results in this dissertation are based on 448 videos that were fed into Emotient and Affectiva for processing, where some videos faced failure due to their video quality.

Reason for selecting the dataset The reason I chose the AM-FED dataset is that it was collected from the internet, capturing spontaneous facial reactions to ads. I believe it's valuable to examine how facial expression analysis systems perform on spontaneous datasets, as this mirrors real-life scenarios.

RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6] represents a comprehensive compilation of facial and audio expressions performed by North American English speakers (professional actors). This database consists of both speeches and songs, featuring a carefully selected "gender-balanced" ensemble of speakers, comprising 12 females and 12 males. These speakers convey seven distinct emotions for speech and five emotions for song performances. The RAVDESS database offers its data in three distinct formats: visual, audio, and audio-visual. For this study, a subset of 720 videos has been chosen from the visual modality, featuring 60 videos each from the 12 speakers (6 male and 6 female).

For the dissertation, 611 videos were selected from 720. The Ravdess results in this dissertation are based on 611 videos that were fed into Emotient and Affectiva for processing, where some videos faced failure due to their video quality.

Reason for selecting the dataset As most of the literature comparing FERs evaluates the performance of the systems on posed and spontaneous expression databases, RAVDESS was introduced to act as the posed expression counterpart of our semi-spontaneous database.

It will be worth checking how well FER systems can perform across a spectrum of facial expressions on Posed datasets.

Data Collection Stage Semi-Spontaneous

The videos of governors were gathered from YouTube using the YouTube downloader.The dataset consists of 60 videos of governors, with the most recent dating back to 2023 and the oldest dating back to 2010. Four to Five videos per Governor were sufficient to be used in the systems. Table 3.1 provides a description of the Governor dataset.

Governor's	Country of Governor	Age	Gender	Ethnicity	Videos Collected
Shaktikanta Das	India	66	Male	South-Asian	4
Dr Reza Baqir	Pakistan	44	Male	South-asian	5
Abdur Rouf Talukder	Bangladesh	58	Male	South-asian	4
Haruhiko Kuroda	Japan	79	Male	East-Asian	5
Yi Gang	China	65	male	East-Asian	5
Christine Lagarde	France	67	Female	White-European	9
Mervyn King	England	75	Male	White-European	5
Andrew Bailey	England	64	Male	White-European	6
Amir Yaron	Israel	59	Male	Semitic	5
Tiff Macklem	Canada	62	Male	White-European	5
Elvira Nabiullina	Russia	60	Female	White-European	5
Janet Yellen	American	77	Female	White-European	5

Table 3.1: About the Collected Data

I have compiled a dataset featuring individuals from various ethnic backgrounds, including South Asian, East Asian, White European, and Semitic. This dataset comprises a total of 12 speakers, with 3 females and 9 males. The speakers were drawn from diverse countries, providing a broad representation. The age range of the participants spans from 40 to 80 years, ensuring a varied demographic. Specifically, the videos were collected for governors from around the world. The source of the video for this study comprises publicly available recordings of speeches, along with its soundtracks. These videos are predominantly shot in controlled settings such as TV studios or professionally captured for broader dissemination, such as media press conferences.

Reason for selecting the dataset I selected YouTube videos featuring politicians and governors for my research. This study draws from public recordings for its video source along with its soundtracks. In most cases, these videos are filmed in well-controlled environments. These performances can be recorded in TV studios or more professionally for distribution like media press conferences. Similarly, governors are professionals who do not fall under the realm of actors following a script, neither a performance nor one to be labeled within the general population. They practiced their speeches and learned how to control or turn up their emotions in order for public display. The governors come from distinct regions composed of many ethnicities with their historical and genetic back-ups. The videos collected are classified as semi-spontaneous. This designation implies that while these professionals are not actors delivering scripted performances, they have practiced their speeches and acquired skills to control or emphasize their emotions for public presentation. This introduces an interesting dynamic for analysis.Given the semi-spontaneous nature of these videos, it becomes essential to

explore how facial expression analysis systems behave when comparing spontaneous expressions with semi-spontaneous ones, as well as contrasting posed expressions with spontaneous ones.

3.4 Facial Emotion Recognition

3.4.1 Architecture of Emotient FACET

Emotient facial expression analysis engine also known as computer expression recognition toolbox(CERT),is the real technology behind the engine used for processing the videos.It automatically codes the 20 different facial action units and 7 different prototypical facial expressions into emotions. [1]. It also estimates the locations of 10 facial features as well as the 3-D orientation (yaw, pitch, roll) of the head.CERT works in the following steps which are described below [1].

1. **Face Detection**: CERT applies the Viola-Jones object detecting algorithm and Boosting algorithms like WaldBoost and GentleBoost [1] are applied for "cascade threshold detection to detect the position of the face of the speaker in the input video for each frame of 33ms. Viola-jones uses Harr-like features, which look at adjacent rectangular regions in each frame and calculate the difference in the sum of the intensity of pixels in those regions. The difference categorizes the frame into subsections of the image. The human face can be detected using Harr features, for example, there is a prominent difference between the intensity of pixels between the eye and cheek region which can be detected by the Harr classifier. The viola-Jones algorithm is applied to each frame, and the face is detected.

2. **Facial Feature Detection**: After face detection, the face window is segmented into facial features like eyes, eyebrows, mouth corners, tip of the node, and lips center. Each facial feature detector, gives us the log-likelihood ratio of that feature being present at that (x, y) location within the face, to being not present at that location. This likelihood term is combined with a feature-specific prior over (x, y) locations within the face to estimate the posterior probability of each feature being present at (x, y) given the image pixels.

3. **Feature Extraction**: The 96x96 pixel patch is "convolved" (Fast Fourier Transform (FFT)) with Gabor filters of different spatial frequencies and orientations.The filter outputs are then fused into a single feature vector for further processing in the later stages.

4. **Action Unit Recognition**: The feature vectors detected are then fed as input to the Support Vector Machine(SVM)to identify different Action Units. The support vector machine is a model to classify data based on a machine learning algorithm. CERT uses a linear support vector machine to classify the feature vector as an action unit separated by a hyperplane.

5. **Intensity of Expressions**: For each action unit, a continuous value (signifying the distance of the input feature vector and the Support Vector Machine's hyperplane) is provided on a frame-by-frame basis. It was observed that the CERT output values were "significantly correlated" with the facial action intensities measured by the FACS experts [19]. This frame-by-frame temporal information on facial action units and their intensities was a resounding success because it was extremely difficult to use manual

coding.

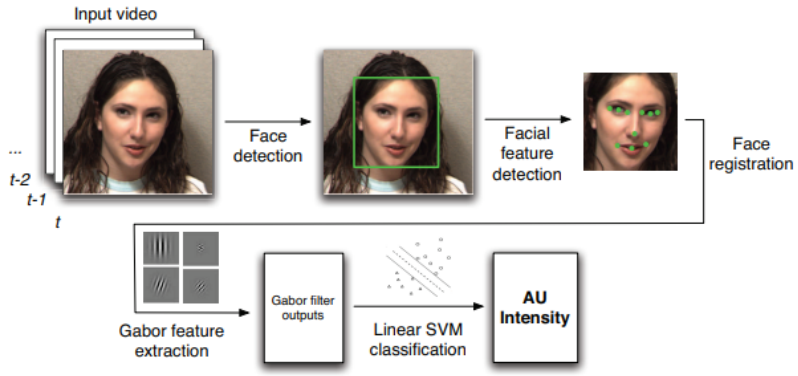

Figure 3.2: Processing pipeline of the Computer Expression Recognition Toolbox (CERT) from video to expression intensity estimates, sourced from [1]

Raw Output of Emotient

Emotient FACET analyses the video on a frame basis.First, the system identifies a face in the image and then annotates 7 facial landmarks for extracting facial features. In this sense, the system selects various AUs also and expressions using various classifiers. The Emotient FACET assigns a score of evidence or proof per emotion. The evidence of the odds ratios in decimal logarithm scale of emotion being present. The positive value indicates that there is more than 50% probability that there is a given emotion whereas the negative value means that it has a lower probability than 50%.This is because the scores can be transformed into probabilities, as shown in the following formula.

$$\text{Prob} = \frac{1}{1 + 10^{-\text{evidence score}}} \tag{3.1}$$

Each row in the data represents the number of detected faces, while the timestamp corresponds to the time interval for face detection in each frame. Specifically, a frame is processed every 33 milliseconds (ms), and the number of detected faces is recorded for each instance

row	timestamp	joy	Joy_Intens	anger	Anger_Inte	surprise	Surprise_I	fear	Fear_Inter	contempt	Contempt_	disgust	Disgust_In	sadness	Sadness_I
3	133.3332	-2.73815	0.001824	-0.20032	0.38669	-4.38659	4.11E-05	-2.13745	0.007234	-0.94556	0.101813	0.675242	0.825606	-0.97228	0.096324
4	166.6665	-2.17021	0.006712	-0.6025	0.199836	-1.45026	0.034246	-0.9792	0.094946	-2.80669	0.001558	-0.5053	0.238033	-2.87334	0.001337
5	199.9999	-2.00072	0.009885	-0.59662	0.202014	-1.18915	0.060761	-0.94539	0.10185	-2.99615	0.001008	-0.62709	0.190937	-2.84006	0.001443
6	233.3332	-1.85239	0.013853	-0.57652	0.209577	-1.27126	0.050826	-0.96274	0.098252	-2.92899	0.001176	-0.48372	0.247162	-2.65824	0.002192
7	266.6665	-2.38054	0.004146	-0.4414	0.265736	-1.5199	0.029321	-0.88227	0.115935	-2.97736	0.001052	-0.00609	0.496496	-2.37574	0.004192
8	299.9999	-2.44193	0.003602	-0.54552	0.221644	-1.25068	0.053161	-0.89792	0.112292	-3.40423	0.000394	-0.24392	0.363165	-2.42784	0.00372
9	333.3332	-2.31889	0.004776	-0.66825	0.176723	-1.14101	0.067403	-0.99745	0.091396	-3.56561	0.000272	-0.54967	0.220003	-2.74399	0.0018
10	366.6665	-1.94122	0.01132	-0.63309	0.188813	-1.4542	0.033947	-0.88687	0.114853	-3.12325	0.000752	0.126102	0.572084	-2.84773	0.001418

Figure 3.3: Emotions Detected by Emotient

Similarly, Emotient detects 20 action units, and the interpretation remains consistent for both negative and positive values

au1	au1_Inten	au2	au2_Inten	au4	au4_Inten	au5	au5_Inten	au6	au6_Inten	au7	au7_Inten	au9	au9_Inten	au10	au10_Inte	au12	au12_Inte	au14	au14_Inte
-0.85572	0.12235	-1.54923	0.027459	0.01006	0.505791	-1.44029	0.035013	0.122167	0.569865	-0.26014	0.354573	-0.4678	0.254047	0.561373	0.784589	-1.50064	0.03061	-1.37723	0.040265
-0.71297	0.162237	-1.02838	0.085651	-0.50206	0.239387	-1.31186	0.0465	-0.66248	0.178665	-0.34663	0.310422	-1.90703	0.012236	0.189344	0.607301	-1.16803	0.063597	-2.28859	0.005119
-0.60814	0.197768	-0.81336	0.133215	-0.5095	0.236284	-1.28108	0.049746	-0.7395	0.154104	-0.46275	0.256256	-1.85342	0.013821	-0.03215	0.481499	-1.15133	0.065925	-2.36794	0.004268
-0.55945	0.21616	-0.78094	0.142072	-0.43462	0.268793	-1.25003	0.053237	-0.67848	0.173323	-0.42609	0.272671	-1.68909	0.02005	0.081008	0.546497	-1.12829	0.069268	-2.32441	0.004716
-0.5601	0.215909	-0.726	0.158202	-0.25876	0.3553	-1.13456	0.068344	-0.74801	0.151569	-0.52491	0.229938	-1.47253	0.03259	0.469672	0.74677	-1.06669	0.07899	-2.25757	0.005496
-0.39745	0.285944	-0.53539	0.225696	-0.19631	0.388878	-1.05992	0.080132	-0.57484	0.210216	-0.38189	0.293318	-1.65137	0.02183	0.226661	0.627593	-1.01078	0.088878	-2.52798	0.002956
-0.38874	0.290059	-0.5338	0.226337	-0.2751	0.346732	-0.99079	0.092678	-0.6041	0.199249	-0.37671	0.295796	-1.84555	0.01407	-0.02671	0.484627	-1.01755	0.087623	-2.61645	0.002413
-0.45823	0.258246	-0.60244	0.19986	-0.12688	0.427478	-1.11695	0.070972	-0.41714	0.276777	-0.19578	0.389169	-1.40988	0.037457	0.365884	0.698987	-0.8928	0.113473	-2.36859	0.004261
-0.35892	0.304396	-0.86879	0.119155	0.136032	0.577672	-1.15762	0.06504	-0.07737	0.455578	-0.03565	0.479491	-1.20366	0.058883	0.826905	0.870346	-0.77358	0.144152	-1.27718	0.050173

Figure 3.4: Action Units Detected by Emotient

3.4.2 Architecture of AFFDEX Affectiva

The entire AFFDEX Affectiva pipeline can be broken down into four steps

1. **Face Detection**:Face detection is performed using the Viola-Jones face detection algorithm [20].34 Landmarks are detected in each facial bounding box.
2. **Feature Extraction**: The facial landmarks define regions of interest in the frame and then the histogram of Oriented features(HOG features) are extracted [21].
3. **Facial Actions**:: Each facial action is then attributed a score between 0 and 100 by a Support Vector Machine (SVM Classifiers).These classifiers were trained on "10000s of manually coded facial images" that were collected from different parts of the world.

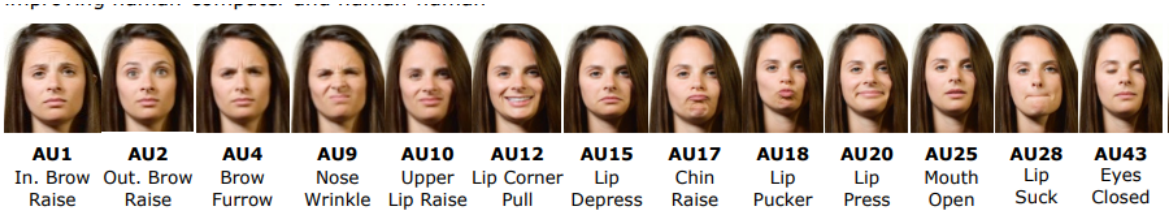

Figure 3.5: AUs detected by the SDK
[22]

4. **Emotion Expressions**: Based on the "combinations of facial actions", the emotions (joy, anger, disgust, sadness, fear, surprise and contempt) are decided. Affectiva uses the Emotional Facial Action Coding System [23] for this purpose. Like their facial action counterparts, scores between 0 and 100 are also assigned to the emotional expressions.

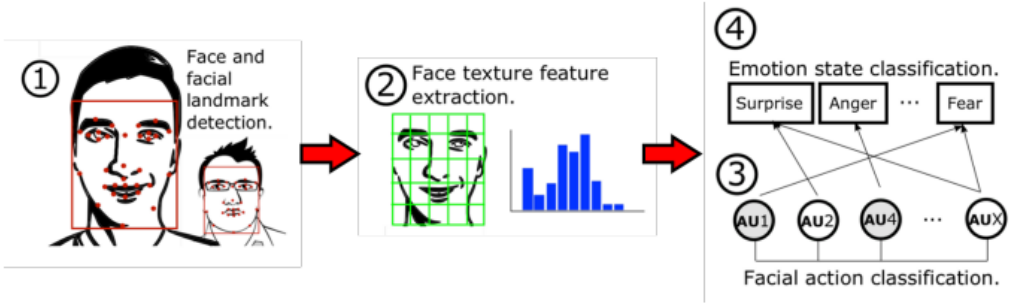

Figure 3.6: Processing pipeline of the AFFDEX
[18]

3.4.3 Raw Output of Affectiva

row	timestamp	anger	contempt	disgust	fear	joy	sadness	surprise
3	133	0.10099	0.107058	0.018976	0.101119	0.065874	0.12486	0.057861
4	166	0.421231	0.065433	0.165667	0.340415	0.136588	0.082823	0.059261
5	199	0.750794	0.058659	0.387996	0.557239	0.160371	0.078655	0.082022
6	233	0.794649	0.064856	0.412069	0.589695	0.137352	0.082093	0.165204
7	266	0.853969	0.075676	0.442475	0.63179	0.108047	0.087514	0.403855
8	299	0.918916	0.089462	0.478038	0.681148	0.083465	0.094158	0.957699
9	333	0.957277	0.098181	0.498196	0.711087	0.071637	0.09797	1.411521
10	366	0.998804	0.108251	0.525147	0.743517	0.062039	0.102175	1.66492

Figure 3.7: Emotions Detected by Affectiva

Affectiva AFFDEX analyses the video on a frame basis.First, the system identifies a face in the image and then annotates 7 facial landmarks for extracting facial features (Anger, Contempt, Disgust, Fear,Joy, Sadness, Surprise).Each row in the data represents the number of detected faces, while the timestamp corresponds to the time interval for face detection in each frame. Specifically, a frame is processed every 33 milliseconds (ms), and the number of detected faces is recorded for each instance.Similarly, Affectiva detects 30 action units(see figure 3.8)

brow furrc	brow raise	cheek raise	chin raise	dimpler	eye closur	eye widen	inner brow	jaw drop	lip corner	lip press	lip pucker	lip stretch	lip suck	lid tighten	mouth ope	nose wrink	smile
0.00614	0.099662	2.401757	0.171383	0	0.001099	0.348642	0.642759	0.005508	3.683871	2.343794	0.000155	0.082436	0.002219	0.009068	0.004127	0.50884	7.815326
0.003896	0.097699	2.336296	0.028955	0	0.028187	0.158222	0.480481	0.934464	0.184665	0.10928	0.000348	0.035182	0.003726	0.008357	72.18871	0.187488	14.38525
0.005461	0.110363	2.307123	0.016718	0	0.047083	0.10298	0.537674	8.726074	0.0784	0.040726	0.000355	0.022447	0.003931	0.010195	99.33596	0.154298	15.83966
0.007417	0.112195	2.208917	0.014652	0	0.01507	0.135794	0.554869	25.4989	0.044717	0.022499	0.000267	0.01882	0.005322	0.009618	99.84501	0.14552	14.4999
0.008978	0.12709	2.0344	0.012652	0	0.01884	0.132839	0.47736	47.07499	0.04479	0.00414	0.00033	0.014954	0.005102	0.008817	99.95276	0.105578	12.4368
0.012446	0.135943	1.878914	0.011408	0	0.013943	0.138216	0.507544	67.88785	0.04201	0.003567	0.000389	0.012953	0.005026	0.009772	99.97929	0.159346	10.20174
0.014727	0.144671	1.630239	0.010381	0	0.008633	0.154644	0.536417	77.25339	0.037781	0.003058	0.000372	0.012612	0.00617	0.009705	99.98713	0.174472	8.958361
0.02129	0.170434	1.670636	0.010299	0	0.002963	0.209668	0.558331	81.17508	0.036141	0.003722	0.000455	0.011137	0.007269	0.010426	99.99065	0.339119	7.648814
0.02247	0.21376	2.100913	0.008065	0	0.001134	0.27602	0.591057	75.14854	0.035716	0.008974	0.000351	0.008344	0.008932	0.008627	99.98188	0.533474	7.562394

Figure 3.8: Action Units Detected by Affectiva

Common Action Units from Both The Systems

Affectiva and Emotient are two facial expression analysis systems that detect different numbers of facial action units (AUs). Specifically, Affectiva detects 30 action units, while Emotient detects 20. In this study, a comparison was conducted and only those action units that are common in both systems were taken into account. The differences between the outputs of the two systems were analyzed at various levels of video description.

Emotient	Affectiva
AU1	Inner Brow Raise
AU6	Cheek Raise
AU4	Brow Furrow
AU7	Lid Tighten
AU9	Nose Wrinkle
AU10	Upper Lip Raise
AU12	Smile
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raise
AU18	Lip Pucker
AU20	Lip Stretch
AU24	Lip Press
AU26	Jaw Drop
AU28	Lip Suck
AU43	Eye Closure

Table 3.2: Common Action Units from Emotient and Affectiva

The focus of previous research by previous students in this domain has primarily centered on the emotions displayed by individuals. However, the current research takes a novel approach by investigating facial subcutaneous muscle activity and its correlation with emotions. This exploration goes beyond the emotions explicitly displayed and explores how facial expressions may be suppressed or exhibited. Table 3.2 illustrates the common action units detected by both systems (Emotient and Affectiva), providing a basis for comparative analysis and insights into the shared facial muscle activities captured by these technologies

3.4.4 A descriptive analysis (Likert Scale)

Flow-Chart for comparing the Likert Scale in two systems (see figure 3.2)

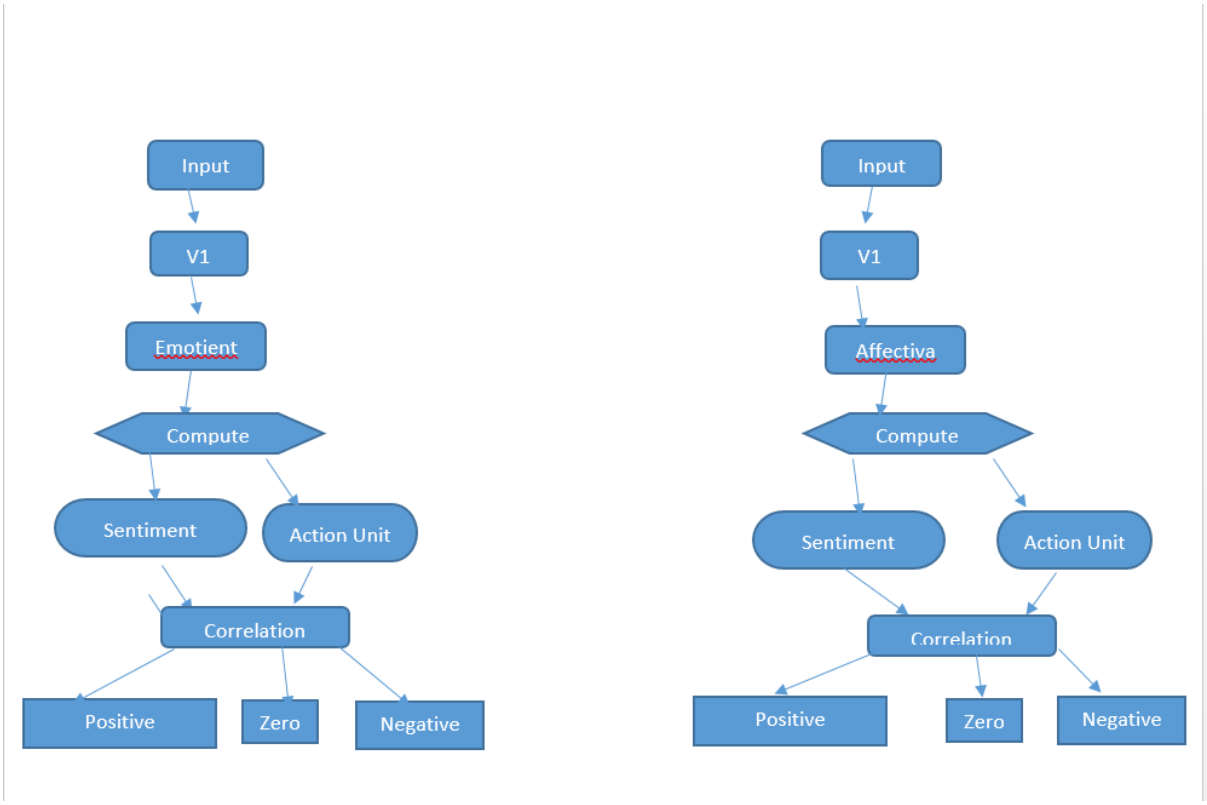

Figure 3.9: Correlations between the inputs

This flow chart shows the correlations between the inputs between two systems, Emotient and Affectiva. For both systems, we have given the same input v1 to Emotient and Affectiva. Thereafter, the systems compute two things, sentiments, and action units, and show the correlation according to the Likert scales [24], whether they are positive, negative, or there is no correlation. [25]

3.5 Statistical Tools and Hypothesis Testing

Statistical testing involves using statistical methods to determine whether there is enough evidence to accept or reject a hypothesis. Before doing statistical testing, it is important to look at the distribution of the variables in question. There are two types of statistical tests: Parametric and Non-Parametric. Parametric tests assume that the distribution of the population from which the samples were taken is a Normal distribution. In contrast, non-parametric tests do not place any constraints on the underlying distribution.

One method of evaluating normalcy involves utilizing a Quantile-Quantile (QQ) plot. A set of visualization techniques called the quantile-quantile plot (QQ plot) proves helpful in confirming the assumption. In this situation, we checked if the data was distributed normally before doing any tests. During testing the data in this study, it was observed that the emotion distributions did not follow a normal distribution. Therefore, the family of non-parametric tests was selected for this study.

3.5.1 Rank order correlation

Spearman's Rank Correlation is a statistical method that doesn't make many assumptions about your data. It helps us understand how two variables relate to each other. While other methods like Pearson's correlation focus on linear relationships, Spearman looks at the order of the data points, not their exact values. In simpler terms, it tells us if the data tends to go up or down together, even if it's not in a straight line [26].

The Spearman correlation coefficient, often denoted as "rho," can range from -1 to 1. If it's -1, it means there's a perfect reverse relationship; if it's 1, it's a perfect direct relationship, and if it's 0, there's no apparent relationship.

This method is especially handy when dealing with data that's not quite numerical, like when you have categories or rankings but don't know if the gaps between them are the same. It's also great when relationships between variables aren't strictly linear, meaning they don't follow a straight line.

In our study, we used Spearman Rank Correlation to figure out how different emotions recognized by facial as well as quality metrics, relate to each other. We categorized these relationships into different levels, following the guidance of Dancey and Reidy's book "Statistics without Maths for Psychology" [27]. In summary, Spearman's Rank Correlation is a versatile tool to understand relationships between variables, especially when you're not sure if the data is linear or when dealing with ordinal data (like rankings).

Kruskal test

The Kruskal-Wallis test is a non-parametric alternative to the ANOVA test, employed to ascertain whether statistically significant differences exist among two or more independent groups [28]. Kruskal Wallis can be invoked to analyze whether a particular value has a significant difference in distribution according to which level of a non-binary discrete category applies. [29].This test focuses on the ranks of the data to determine if there are significant differences among the groups.The null hypothesis H0 in the Kruskal-Wallis H-test is that the two distributions are similar or the samples of the two sets of observations originate from the same distribution. It usually compares two independent groupsThe alternate hypothesis H1 is that the two distributions are statistically different. The null hypothesis is tested with the help of a metric called the p-value which is the probability of an event occurring given that the null hypothesis is true. Now, if the p-value is less than a threshold (0.05) the null hypothesis is rejected in favour of the alternate hypothesis .

Wilcoxon Test

The Wilcoxon rank-sum test is a non-parametric method employed to determine whether there are significant differences between two independent groups. This test assesses whether the distributions of the two groups differ significantly by comparing the ranks of the observations from both groups. The test yields a p-value that helps to decide whether to reject the null hypothesis. For example, It can be useful in understanding whether a single system differs in some measure according to a binary factor, such as biological sex is within the data that we have at hand [30]. In situations where the Wilcoxon rank-sum test indicates significant differences between groups, a pairwise Wilcoxon rank-sum test can be conducted.

This subsequent test enables the identification of specific pairs of groups that exhibit significant differences. It involves performing individual Wilcoxon rank-sum tests for each pair of groups and then adjusting the p-values to control for multiple comparisons using methods like the Bonferroni correction [31].

3.5.2 Contingency Table and Chi-Squared Test

Contingency tables are used to show the relationship between two or more categorical variables [32]. It shows how often these categorical variables occur together [33]. Each cell in the table represents the count or frequency of observations that fall into a particular combination of categories. Contingency tables are commonly employed to investigate associations, dependencies, or patterns among categorical variables. A chi-square contingency table, also known as a chi-squared test of independence, is a statistical test used to determine whether there is a significant association between two categorical variables in a contingency table [34]. The test is based on the chi-square statistic, which quantifies the difference between the observed frequencies in the table and the frequencies that would be expected if the variables were independent. Standard residuals are values calculated from the observed and expected frequencies in a contingency table.

3.6 Conclusion

There are many systems available but two systems (Emotient(FACET)) and Affectiva(AFFDX) were chosen. Emotient was amongst the first developed, closely followed by (AFFDX). Emotient and Affectiva were used because they have been well tested. Other software, such as Microsoft Azure https://azure.microsoft.com/en-us) were rejected due to licensing difficulties.

4 Results

4.1 Hypothesis Testing

In this dissertation, I argue that the emotion-performance of the two recognition systems is similar, and therefore there is no different training or test or architectural bias. Furthermore, a sub-hypothesis has been established assuming that emotion recognition should not be affected by factors such as ethnicity, age, and sex. Numerous statistical tests have been conducted to verify whether the null hypothesis holds or whether the alternate hypothesis takes precedence, as elucidated in chapter 3

Methods And Data Selection

A collection of videos, featuring both posed and spontaneous expressions, was individually presented to both Emotient and Affectiva. These systems provided information about the activation values of specific units and the likelihood of individuals exhibiting particular emotions. To evaluate the degree of agreement between the two systems, the order of activation units and emotion likelihood was computed for each frame. The relationship between the sequences of the two outputs was then examined to estimate whether the systems agreed (positive correlation), disagreed (negative correlation), or had no correlation (zero correlation) with each other.

To evaluate the correlation between the outputs of Affectiva AFFDEX and Emotient FACET, a Likert scale was employed, drawing inspiration from the psychological tool outlined by Batterton [35]. The Likert scale utilized a three-point continuum, classifying correlations into negative, positive, and zero categories.

Guided by the principles presented in Batterton's paper [35], a technique was devised to create a bin range from -1 to +1. This binning strategy facilitated the categorization of correlation values: those between -1 and -0.1 were considered strongly negative, values around 0 indicated no correlation, and values between 0.1 and 1 were regarded as strongly positive correlations.

This technique, inspired by the Likert scale and adapted from Batterton's insights, provided a nuanced approach to describe the correlation observed between the outputs of Affectiva AFFDEX and Emotient FACET in the context of the research. It served as a valuable tool to systematically categorize and interpret the degree and nature of correlation, contributing to a more detailed analysis of the systems' agreement or disagreement

Following the application of the Spearman test, which checks for associations between the outputs of the two systems, a Likert scale was employed to categorize the observed correlations. The three-point Likert scale was particularly useful in distinguishing whether the correlations

were positive, negative, or zero.

4.2 Our Test Corpus

We have utilized a total of 1157 videos, which comprise the gold-standard data and our collection of political and business leaders.

Gold Standard Database

1. Ravdess [6]
2. AM-FED+ [5]

Our Collection

1. State Bank Governors
2. Political Leaders

Gold-Standard			Total Videos	Average Duration	Max	Min
	Posed	Ravdess	611	00:04	00:05	00:02
	Spontaneous	Am-Fed	448	12:45	51:13	00:32
Our Collection	Spontaneous					
		State Bank Governor's	59	03:00	04:20	01:06
		Political Leaders	39	03:10	6:40	3:10

Table 4.1: Our Test Corpus

4.3 Likert-Scale Description of Similarity/Differences in the outputs of Affectiva and Emotient

Three separate tables and one combined table were created to correspond to different datasets: posed datasets (Ravdess), semi-spontaneous datasets (Governors from my collection), and spontaneous datasets (Am-fed+). A three-point Likert scale was used to note down the values falling within the range of negative, zero, and positive correlations for each dataset.

4.3.1 Emotion-Level Agreement

The first table is a combined summary, focusing on emotion levels, to examine differences in the outputs of Affectiva and Emotient. This table provides an overall view of how the systems compare in terms of emotional expression. I conducted a detailed analysis focusing on the distribution of data points across different emotions. The table below represents a combined summary of all datasets, including posed, semi-spontaneous, and spontaneous expressions. The data is categorized into three groups: Negative, Zero, and Positive, with each emotion assigned a specific count.

Emotion	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise	Total all sentiments
Negative	413	305	200	188	30	223	154	1513
Zero	124	133	113	116	43	115	114	758
Positive	620	719	844	853	1084	819	889	5828
Total Points	1157	1157	1157	1157	1157	1157	1157	8099

Table 4.2: Combined Table for all Data for Emotions

Table 4.2 provides an overview of how many data points were recorded for each emotion in the negative, zero, and positive categories. For instance, under 'Anger,' there were 413 data points in the negative category, 124 in the zero category, and 620 in the positive category. A notable observation is that 'Joy' exhibits the highest number of data points in the positive category. Specifically, out of the 1157 total data points, 1084 were recorded in the positive correlation. This indicates that 'Joy' was the most prominently expressed emotion across the videos, with a substantial 93.5% of occurrences falling into the positive correlation.

The 'Total Data Points' row in the table signifies the cumulative count of all data points for each emotion, remaining constant at 1157. This cumulative total offers a comprehensive perspective on the prevalence of each emotion, highlighting their distribution across negative, zero, and positive correlations.

Emotion	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise
Negative	36%	26%	17%	16%	3%	19%	13%
Zero	11%	11%	10%	10%	4%	10%	10%
Positive	54%	62%	73%	74%	94%	71%	77%

Table 4.3: Overall Emotion Percentages

After analyzing Table 4.2, which provided a detailed breakdown of data points for each emotion, we observed a significant prominence of the 'Joy' emotion, particularly in the positive category. To further illustrate this observation, Table 4.3 presents the overall percentages of each emotion across negative, zero, and positive correlations. Table 4.3 offers a condensed view of the prevalence of each emotion across all datasets. Notably, **'Joy'** stands out with a substantial **94%** occurrence in the positive category, indicating that this emotion was prominently expressed in the majority of the videos. In contrast, **'Anger'** exhibits a lower percentage of **54%**, suggesting that it was less prevalent compared to **'Joy'** in the analyzed videos. This overview provides valuable insights into the distribution of emotions, shedding light on the varying degrees of expression captured by the system in different emotional categories.

Emotion	Average	Standard Deviation	Min	Max	Maxvariance	Overall Detection
Negative	216.14	119.90	30	413	1.642	19%
Zero	108.29	29.66	43	133	0.833	9%
Positive	832.57	144.46	620	1084	1.740	72%

Table 4.4: Descriptive Statistics for Emotions

On average, emotions in the positive category have more data points, with an average of 832.57. The negative category has an average of 216.14, and the zero category has the lowest

average of 108.29. At least 30 data points were recorded for each emotion in the negative category. In the zero category, the minimum is 43, and in the positive category, the minimum is 620, which emphasizes the prevalence of positive emotions in the dataset. Conversely, the maximum values demonstrate the highest count of data points for each emotion. The maximum value in the negative category is 413, in the zero category it is 133, and in the positive category it is 1084, indicating a substantial range of expressions captured. A maximum variance value of 1.642 suggests that people in the videos show a significant range of different negative emotions on their faces. The maximum variance value of 1.740 in the positive category proposes that people in the videos exhibit a wide range of variations when expressing positive emotions on their faces. It means that the detection of different moods on average is different.When analyzing all the data from three sources, the system found that **72%** of the emotions were **positive**. This means that, **on average**, the system were capable of identifying **positive emotions**. There was a high positive correlation between the systems and different emotions when people expressed themselves positively.

4.3.2 Activation-Level Agreement

After reviewing table 4.4 for emotion level agreement , I went into more detail. I focused on how the data points were distributed across various action unit categories associated with different emotions. Table 4.5 combines information from various datasets, including posed, semi-spontaneous, and spontaneous expressions. The data is divided into three groups: Negative, Zero, and Positive, and it tells us how many times each action unit was identified. The table provides combined data for 15 action units that were common in both systems, Emotient and Affectiva, and measures the accuracy of which action units were the highest positively correlated and which action units were the least negatively correlated.The 'Total Data Points' row in the table signifies the cumulative count of all data points for each emotion, remaining constant at 1157 as already explained in section **4.3.1**.

Action Unit	Negative	Positive	Zero	Total All Points
AU1	206	850	101	1157
AU6	144	916	97	1157
AU4	235	809	113	1157
AU7	132	931	94	1157
AU9	149	910	98	1157
AU10	137	925	95	1157
AU12	41	1080	36	1157
AU15	196	734	227	1157
AU17	119	934	104	1157
AU18	158	885	114	1157
AU20	436	565	156	1157
AU24	8	1021	128	1157
AU26	108	968	81	1157
AU28	200	841	116	1157
AU43	150	889	118	1157
Total all Action Units	2419	13258	1678	17355

Table 4.5: Combined Table for All Data for Action Units

Overall Action Units Percentages

After analyzing Table 4.5, which provided a detailed breakdown of data points for each action unit, we observed a significant prominence of the **' AU12'** situated with emotion **smile** , particularly in the positive category Under **'AU12,'** there were only 41 data points in the negative category, 36 in the zero category, and 1080 in the positive category. A notable observation is that **'AU12'** exhibits the highest number of data points in the positive category for 93% of occurrences among the total 1157 data points. On the other hand, **'AU20' (lip stretch)** was observed the least in the positive correlation, indicating the minimal presence in the videos. This implies that **'AU20' (lip stretch)** was less frequently expressed in a positive emotional context across the analyzed videos there is not much difference in negative and positive categories which means **'AU20'** is the worst among all the action units that was found very less positively correlated in the videos. The low percentage count in the positive category suggests that this specific action unit may not be as strongly associated with positive emotions or expressions in the observed dataset.

Action Unit	Negative	Zero	Positive
AU1	18%	9%	73%
AU6	12%	8%	79%
AU4	20%	10%	70%
AU7	11%	8%	80%
AU9	13%	8%	79%
AU10	12%	8%	80%
AU12	4%	3%	93%
AU15	17%	20%	63%
AU17	10%	9%	81%
AU18	14%	10%	76%
AU20	38%	13%	49%
AU24	1%	11%	88%
AU26	9%	7%	84%
AU28	17%	10%	73%
AU43	13%	10%	77%

Table 4.6: Overall Action Units Agreement

Going deeper into the study, we found some interesting things about how people express emotions in videos. One interesting point is that when people are **happy or smiling (connected with 'AU12')**, this expression was seen a lot, about **93%** of the time when the system detected positive feelings.

Another finding was about **'AU24,'** related to **lip pressing**. It appeared quite often, suggesting a specific facial expression connected with certain emotions in the videos.

We also noticed **'AU7' and 'AU10,'** which were connected to intense facial expressions and **anger**, respectively. These action units were found **positively correlated** about **80%** of the time. This connects with what we saw in the **emotion-level-agreement(4.3.1)**, where **anger** didn't have strong agreement overall.

So, even though **anger** didn't show up much in the general **emotion-level-agreement(4.3.1)**, when it did, it was often linked with specific facial actions. This tells us more about how people

express themselves, adding a layer of detail to our study.

	Average	Standard deviation	Min	Max	Minvariance	Maxvariance	Overall Detection
Negative	161.26	96.44	8	436	-1.58	2.84	14%
Zero	111.21	40.91	36	227	-1.85	2.81	10%
Positive	883.91	121.08	565	1080	-2.63	1.61	76%

Table 4.7: Overall Descriptive Statistics for Action Units for all Data

When analyzing all the data from three sources (Ravdess, AM-FED+, Governors), the system found that **76%** of the **action units** were **positively correlated**. This means that, **on average**, the system was capable of identifying **positive emotions**. There was a high positive correlation between the systems and different action units when people expressed themselves positively **as aready explained in table 4.4 descriptive statistics for emotions**.

4.3.3 Variance across Posed ,Spontaneous and Semi-Spontaneous Data Set

4.2.3.1 Emotion-Level Differences

In our study, we compared how well the systems detected emotions in different datasets. We used a three-point scale **as already explained in methods 4.1** - negative, zero, and positive - to measure their accuracy. The table below summarizes the overall results for three datasets: RAVDESS, AM-FED, and Politicians/Governors Datasets.

Datsets	Negative	Zero	Positive
RAVDESS	19%	5%	76%
AM-FED	20%	15%	65%
Our-Collection	12%	11%	76%

Table 4.8: Emotion Level Percentage Agreement

The numbers represent the percentage of accurate emotion detection. Interestingly, we found that **RAVDESS (posed dataset)** and **our collection Semi-Spontaneous dataset** showed a higher accuracy of **76% positive detection**, while the **AM-FED (spontaneous) dataset** had a slightly lower accuracy at **65%**. This difference might be because **AM-FED** videos have more **visual distractions** in the videos, making it challenging for the systems to accurately detect emotions(See table 4.8).

In our investigation, we delved into the positive correlations across various emotions in three datasets: RAVDESS, AM-FED, and Semi-Spontaneous. The table below showcases the percentage accuracy of detecting each emotion in the **positive category**.

Across Emotion	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise
Ravdess	47%	72%	83%	82%	97%	70%	83%
Am-Fed	62%	49%	62%	61%	89%	69%	65%
Our-Collection	60%	59%	60%	79%	99%	85%	93%

Table 4.9: Positive Correlation across all emotion

Notably, both **RAVDESS and our collection dataset** demonstrated high accuracy in detecting **joy**, reaching **97% and 99%**, respectively. On the other hand, the **AM-FED** dataset showed a slightly **lower** accuracy for **joy**, recording **89%**. Interestingly, **surprise emotion** were most accurately detected in the **Semi-Spontaneous dataset (our collection)**, with a remarkable **93%**. This could be attributed to the controlled nature of expressions by politicians and governors, leading to more consistent emotion detection. Conversely, **anger** was less frequently detected across all three datasets compared to other emotions.(see table 4.9)

Interaction between the Likert Scale x Emotions

The **'Likert Scale'** indicates whether the correlation is **negative, positive, or zero** The following tables for **Likert Scales x Emotion** provide an insightful view of the interaction between emotions across three distinct categories: negative, zero, and positive, for each dataset. These tables are presented on a 3-point Likert scale, offering a nuanced perspective on the distribution of emotions.We will now examine the distribution of positive, negative, and no agreement for each data set.(Table 4.10, 4.11, and 4.12) show the data for emotions in terms of 3-point Likert scales differently. In the analysis of emotional expressions within the **RAVDESS dataset**,**AM-FED+ dataset**, and **Our-Collection dataset** we examined the distribution of **sentiments** across three categories: **negative, zero, and positive**. The table below provides a detailed breakdown for each emotion for each dataset, showcasing the count of data points in each sentiment category and the overall total.

Table for Likert Scale x Emotion for Posed(Ravdess dataset)

Emotion	Negative	Zero	Positive	Total All points
Anger	293	33	285	611
Contempt	130	41	440	611
Disgust	77	26	508	611
Fear	72	35	504	611
Joy	12	9	590	611
Sadness	143	41	427	611
Surprise	68	36	507	611
Total all Sentiments	795	221	3261	4277

Table 4.10: Table for Ravdess data

The following table(4.10) meticulously breaks down the count of data points for each emotion across different sentiment categories—negative, zero, and positive. The 'Total all Sentiments' row quantifies the distribution of positive, negative, and zero values, offering a comprehensive snapshot of the emotional expressions captured within the dataset. In this context, the total represents the sum of all sentiments, providing valuable insights into the prevalence of different emotional states across the Ravdess dataset, totaling 4277 data points.

Table for Likert Scale x Emotion for Spontaneous(AM-FED+) dataset

Emotion	Negative	Zero	Positive	Total all Points
Anger	95	77	276	448
Contempt	154	73	221	448
Disgust	103	68	277	448
Fear	106	70	272	448
Joy	18	33	397	448
Sadness	75	64	309	448
Surprise	82	75	291	448
Total all Sentiments	633	460	2043	3136

Table 4.11: Table for AM-FED+ data

The following table(4.11) meticulously breaks down the count of data points for each emotion across different sentiment categories—negative, zero, and positive. The 'Total all Sentiments' row quantifies the distribution of positive, negative, and zero values, offering a comprehensive snapshot of the emotional expressions captured within the dataset. In this context, the total represents the sum of all sentiments, providing valuable insights into the prevalence of different emotional states across the AM-FED+ dataset, totaling 3136 data points.

Table for Likert Scale x Emotion for Semi-Spontaneous(Our-Collection) dataset

Emotion	Negative	Zero	Positive	Total all Points
Anger	25	14	59	98
Contempt	21	19	58	98
Disgust	20	19	59	98
Fear	10	11	77	98
Joy	0	1	97	98
Sadness	5	10	83	98
Surprise	4	3	91	98
Total all sentiments	85	77	524	686

Table 4.12: Table for Our-Collection Data

The following table(4.12) breaks down the count of data points for each emotion across different sentiment categories—negative, zero, and positive. The 'Total all Sentiments' row quantifies the distribution of positive, negative, and zero values, offering a comprehensive snapshot of the emotional expressions captured within the dataset. In this context, the total represents the sum of all sentiments, providing valuable insights into the prevalence of different emotional states across the governors dataset, totaling 686 data points.

4.2.3.2 Activation -Level Differences

In our study on **emotion level as explained in section 4.2.3.1**, we explored how well the systems performed in recognizing emotions across three datasets: RAVDESS (posed dataset), AM-FED (spontaneous dataset), and our collection (Semi-Spontaneous dataset). Surprisingly,

while **RAVDESS and our Semi-Spontaneous** collection showed a solid **76%** accuracy in detecting **positive emotions** in the **emotion level agreement**, the **activation level agreement** brought out some interesting differences. **RAVDESS** displayed a notable **81%** accuracy in detecting **positive emotions** at the activation level. On the other hand, **AM-FED+** had a slightly lower accuracy at **69%**, and our collection stood out with an impressive **84%** accuracy. This suggests that when we dive deeper into the details of activation levels , the systems tend to perform even better, especially in **Our-Collection Semi-Spontaneous dataset**, which showed **higher accuracy 84%** in detecting positive categories in **activation level agreement** compared to **emotion level agreement**(See Table 4.13).

Datasets	Negative	Zero	Positive
RAVDESS	14%	5%	81%
AM-FED	15%	16%	69%
Our-Collection	8%	8%	84%

Table 4.13: Activation level Percentage Agreement

Earlier, we scrutinized the emotion levels to discern overall accuracy, and now our attention shifts to the activation level, aiming to identify which action units exhibit high positive correlation with emotions in each dataset. This approach allows us to unravel the intricacies of emotion detection, offering a more nuanced perspective on how specific facial action units contribute to the recognition of positive emotional states. The results are presented in the subsequent table.

Action Unit	Ravdess	Am-fed	Semi-Spontaneous
AU1	80%	63%	85%
AU6	83%	71%	94%
AU4	72%	65%	81%
AU7	86%	71%	89%
AU9	82%	71%	89%
AU10	85%	72%	83%
AU12	96%	89%	99%
AU15	75%	44%	80%
AU17	91%	67%	81%
AU18	79%	72%	80%
AU20	40%	61%	52%
AU24	99%	73%	92%
AU26	89%	76%	85%
AU28	77%	66%	77%
AU43	78%	71%	93%

Table 4.14: Positive Correlation across all Action Units

One standout was **'AU12'**, associated with smiles or happiness. Both RAVDESS and our collection showed high accuracy, reaching **96% and 99%,** respectively. However, AM-FED had a slightly lower accuracy at **89%**. Another noteworthy finding was for **'AU6,'** associated with **cheek raising**. The Semi-Spontaneous dataset (our collection) demonstrated

the highest accuracy at **94%,** while RAVDESS and AM-FED had accuracies of **83% and 71%,** respectively.

Surprisingly, **'AU43,'** linked to **eye closure**, showed high accuracy in **our collection** at **93%,** compared to RAVDESS and AM-FED with **78% and 71%**. This suggests that some videos in our collection prominently featured eye closure expressions.

These findings highlight the nuanced variations in action unit detection, with certain action units being more accurately identified in specific datasets. It's intriguing to note how different datasets influence the detection of facial expressions associated with various action units (See Table 4.14). Additionally, it's important to note that we focused solely on **positive correlation** values to provide a clearer picture of action units that are highly positively correlated with emotions in each dataset.

Dominant Action Units for Each Emotion

Some action units play a crucial role in expressing specific emotions, as suggested by various authors(see table 4.15). In the context of **joy**, the presence of **AU12(smile)** and **AU6 (Cheek raise)** is often emphasized. When it comes to expressing **anger**, authors highlight the significance of **AU4 (Brow Furrow)**, **AU7 (Lid Tighten)**, **AU10 Upper Lip Raise**, **AU24 (Lip Press)**. Additionally, for expressing surprise, **AU1 (Inner Brow raise)** is often recommended. These associations are detailed in Table 4.15, providing insights into which action units are commonly linked to the expression of specific emotions.

Emotions	Anger	Joy	Sadness	Surprise	Disgust	Fearful	Authors
Action Units	AU4 AU7 AU10 AU 17	AU12 AU25 AU 6	AU 4 AU17 AU 1 AU 6 AU 11	AU 1 AU 2 AU 25 AU 26 AU 5	AU 9 AU 10 AU 17 AU 4 AU 24	AU 1 AU 4 AU 20 AU 25 AU 5 AU2 AU26	Du,Tao, Martinez [36]
Action Units	AU23 AU7 AU17 AU4 AU2	AU12 AU6 AU26 AU10 AU23	AU15 AU1 AU4 AU17 AU10	AU27 AU2 AU1 AU5 AU26	AU9 AU7 AU4 AU17 AU6	AU20 AU4 AU1 AU5 AU7	S. Velusamy [9]
Action Units	AU23 AU24	AU 12	AU 1 AU 4 AU 15 AU11 AU 6 AU 5	AU 1 AU 2 AU 5	AU 9 AU 10	AU 1 AU 2 AU 4 AU 5	LUCEY [37]

Table 4.15: Dominant Action Units for Each Emotions

Interaction between the Likert Scale x Action Units

The tables regarding **Likert Scales x Action Units** offer valuable insights into how action units interact across three main categories: negative, zero, and positive, for each dataset. These tables, based on a 3-point Likert scale, provide a detailed perspective on how action units are distributed. To better understand this, we'll explore the distribution of positive, negative, and no agreement for each dataset in terms of activation level, as explained in **Section 4.2.3.1**, which focuses on the interaction between the Likert scale and emotions. (Table 4.16, 4.17, and 4.18) show the data for action units in terms of 3-point Likert scales differently.

Table for Likert Scale x Action Units for Posed (Ravdess) dataset

Action Unit	Negative	Zero	Positive	Total all points
AU1	96	28	487	611
AU6	67	38	506	611
AU4	141	29	441	611
AU7	59	25	527	611
AU9	28	28	504	611
AU10	64	27	520	611
AU12	16	11	584	611
AU15	116	36	459	611
AU17	20	35	556	611
AU18	92	34	485	611
AU20	302	67	242	611
AU24	5	2	604	611
AU26	42	23	546	611
AU28	107	34	470	611
AU43	80	52	479	611
Total all ActionUnits	1286	469	7410	9165

Table 4.16: Table for Ravdess for Action Units

The following table(4.16) meticulously breaks down the count of data points for each action units across different categories—negative, zero, and positive. The 'Total all ActionUnits' row quantifies the distribution of positive, negative, and zero values, offering a comprehensive snapshot of the activation level expressions captured within the dataset. In this context, the total represents the sum of all points, providing valuable insights into the prevalence of different action unit states across the Ravdess dataset, totaling 9165 data points.

Table for Likert Scale x Action Unit for Spontaneous(AM-FED+) dataset

Action Unit	Negative	Zero	Positive	Total all Points
AU1	105	63	280	448
AU6	74	56	318	448
AU4	86	73	289	448
AU7	69	62	317	448
AU9	64	65	319	448
AU10	62	62	324	448
AU12	25	24	399	448
AU15	68	183	197	448
AU17	89	60	299	448
AU18	59	67	322	448
AU20	106	70	272	448
AU24	0	121	327	448
AU26	58	51	339	448
AU28	86	66	296	448
AU43	67	62	319	448
Total all Actionunits	1018	1085	4617	6720

Table 4.17: Table for AM-FED+ for Action Units

The following table(4.17) meticulously breaks down the count of data points for each action unit across different categories—negative, zero, and positive. The 'Total all ActionUnits' row quantifies the distribution of positive, negative, and zero values, offering a comprehensive snapshot of the activation level expressions captured within the dataset. In this context, the total represents the sum of all points, providing valuable insights into the prevalence of different action unit states across the AM-FED+ dataset, totaling 6720 data points.

Table for Likert Scale x Action Units for Semi-Spontaneous(Our-Collection) dataset

Action Unit	Negative	Zero	Positive	Total all points
AU1	5	10	83	98
AU6	3	3	92	98
AU4	8	11	79	98
AU7	4	7	87	98
AU9	6	5	87	98
AU10	11	6	81	98
AU12	0	1	97	98
AU15	12	8	78	98
AU17	10	9	79	98
AU18	7	13	78	98
AU20	28	19	51	98
AU24	3	5	90	98
AU26	8	7	83	98
AU28	7	16	75	98
AU43	3	4	91	98
Total all Action Units	115	124	1232	1470

Table 4.18: Table for Our-Collection Data for Action Units

The following table(4.18) meticulously breaks down the count of data points for each action unit across different categories—negative, zero, and positive. The 'Total all ActionUnits' row quantifies the distribution of positive, negative, and zero values, offering a comprehensive snapshot of the activation level expressions captured within the dataset. In this context, the total represents the sum of all points, providing valuable insights into the prevalence of different action unit states across The Semi-spontaneous (Our-Collection) dataset, totaling 1470 data points.

Performance on Emotions vs Action Units

In our analysis, we examined how well two systems, Affectiva and Emotient, performed in detecting specific emotions, focusing on the example of **anger** and **joy**. For each **action unit** associated with **anger** and **joy**, we noted the **highest correlation values** and **lowest correlation values from** found in a random video from our dataset.

Tables (4.19 and 4.20) represent the correlation for anger and the correlation for joy with action units. The white box denotes Affectiva results, while the green box represents Emotient outcomes. These findings contribute to our understanding of how well these systems align with existing research on facial expressions and emotions.

Anger

Figure 4.1: Anger Emotion from Emotient and Affectiva

Activation AU for Anger	Lowest Correlation	Highest Correlation	Authors
AU4	-0.02	0.58	S. Velusamy ([36], [9])
AU7	0.35	0.19	S. Velusamy ([36], [9])
AU10	0.32	0.33	S. Du ([36])
AU17	0	0.2	S. Velusamy ([9])
AU24	0.42	0.32	Lucey ([37])

Table 4.19: Correlation for Anger with Action Units

Upon examining the highest and lowest correlations for the emotion of anger in the presented table, a stark contrast emerges between the Emotient and Affectiva systems. For the specific action unit AU4 (Brow Furrow), the Lowest correlation registers a minimum correlation of only -0.02, indicating a very weak or even inverse relationship. In stark contrast, for that particular action unit highest correlation, reached a substantial 58%. This substantial disparity suggests notable differences in how these systems interpret and correlate action units with the emotion of anger.(see table 4.19).

Joy

Figure 4.2: Joy Emotion from Emotient and Affectiva

Activation AU for Joy	Lowest Correlation	Highest Correlation	Authors
AU6	0.02	0.61	S.velusamy [36] [9]
AU12	-0.03	0.74	Lucey, S.velusamy [37] [36] [
AU26	0.33	0.14	S.velusamy [9]
AU10	-0.2	0.53	S.velusamy [9]

Table 4.20: Correlation for Joy with Action Units

Upon examining the highest and lowest correlations for the emotion of joy in the presented table, a stark contrast emerges between the Emotient and Affectiva systems. For the specific action unit AU12 (Smile), the Lowest correlation registers a minimum correlation of only -0.03, indicating a very weak or even inverse relationship. In stark contrast, for that particular action unit highest correlation, reached a substantial 74%. This substantial disparity suggests notable differences in how these systems interpret and correlate action units with the emotion of joy.(see table 4.20).

Results of Correlation Between Variables

To assess the relationships and interactions between different variables, we established a three-category Likert scale (negative, positive, zero), considered the source of datasets (posed, spontaneous or semi-spontaneous), and examined various emotions (7 in total). Using Python, we generated a contingency table to observe potential associations between these variables.

In this analysis, we calculated the degrees of freedom, p-value, and chi-square to determine the statistical significance of observed interactions. The purpose was to ascertain whether there

are meaningful connections between Likert scale ratings, the source of datasets, and different emotional expressions as well as activation level. These statistical measures aid in deciding whether to accept or reject hypotheses regarding the relationships among these variables.

4.3.4 Emotion Level Correlation

In this analysis, a comprehensive examination of the relationships between the source of datasets (RAVDESS AM-FED+ Our-Collection), the Likert scale categories (negative, positive, and zero), and various emotions (Anger, Contempt, Disgust, Fear, Joy, Sadness, Surprise) was conducted. The data were organized into a contingency table, which facilitated the exploration of potential interactions between these variables.

The contingency table was structured to showcase the distribution of sentiment categories (negative, positive, zero) for each emotion across the different datasets. This enabled a closer look at how emotions were expressed in **posed (RAVDESS)**, **spontaneous (AM-FED+)** , and **semi-spontaneous(Our-Collection)** scenarios, categorized by the Likert scale , recall that tables **as explained in the section 4.2.3.1(Table 4.10, 4.11, and 4.12)** for each dataset,

The subsequent step involved conducting a **Chi-square test**, which assessed whether there were significant associations between the **Likert scale** categories **(Negative , Positive or Zero)**, **Source** datasets(Ravdess, AMF-ED+, Our-Collection), and **Emotions**. The Chi-square test produced a statistic, a **p-value**, and the **degree of freedom**. The statistic quantified the difference between the observed and expected frequencies in the contingency table. The **p-value** indicated the probability of obtaining such results by chance, and the **degree of freedom** reflected the flexibility in the distribution of variables. The significance of the Chi-square test is determined by the p-value. If the p-value is less than 0.05, it indicates that there is a statistically significant association between the Likert scale categories, source datasets, and emotions. In this case, we would reject the null hypothesis [38], suggesting that the observed distribution in the contingency table is not likely due to random chance. On the other hand, if the p-value is greater than 0.05, we do not have enough evidence to reject the null hypothesis, implying that the observed distribution may occur by chance.

This analytical approach allowed for a robust exploration of how sentiments and emotions varied across different datasets and Likert scale categories. It offered valuable insights into the interplay between the three datasets, contributing to a comprehensive understanding of emotion detection in diverse scenarios.

4.3.4.1 Ravdess vs AM-FED+

In consideration of the data presented in **Table 4.10 and 4.11** for Ravdess and AM-FED+, I conducted a comprehensive analysis to calculate the correlation between variables.

Interaction	Degree of freedom	p-value	chi-square	Reject or Accept
Likert Scale x Emotion	12	0	543.34	Rejected
Likert Scale x Source	2	0	211.34	Rejected
Source x Emotion	6	1	0	Not Rejected

Table 4.21: Correlation of Emotional Outputs for Ravdess vs AM-FED+

1. **Likert Scale x Emotion**

   The Chi-square statistic of 543.34 with a p-value close to zero (1.3e-108) indicates a highly significant association between Likert scale categories and emotions. The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 12 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables. Certainly! The table represents the observed counts for each combination of the Likert Scale and Emotion. Each row corresponds to a Likert Scale category (Negative, Positive, or Zero), and each column corresponds to a specific emotion (Anger, Contempt, Disgust, Fear, Joy, Sadness, Surprise). The numbers in the table indicate how many data points fall into each combination. For instance, in the "Negative" Likert Scale category, there are 388 occurrences of Anger, 284 of Contempt, 180 of Disgust, and so on.This breakdown provides a detailed view of the distribution of emotions across different Likert Scale responses. The observed counts presented in the table make a total of 1059 videos considered, encompassing data from both the RAVDESS and AM-FED+ datasets.

Likert Scale	**Anger**	**Contempt**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**
Negative	388	284	180	178	30	218	150
Positive	561	661	785	776	987	736	798
Zero	110	114	94	105	42	105	111
Total	1059	1059	1059	1059	1059	1059	1059

Table 4.22: Observed counts for each combination of Likert Scale and Emotion

Likert Scale	**Anger**	**Contempt**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**
Negative	37%	27%	17%	17%	3%	20%	14%
Positive	53%	62%	74%	73%	93%	70%	75%
Zero	10%	11%	9%	10%	4%	10%	11%

Table 4.23: Percentage agreement between RAVDESS vs AM-FEd+

The table 4.23 illustrates the percentage agreement between the Likert Scale categories "Negative," "Positive," and "Zero" for each emotion when comparing the **RAVDESS** and **AM-FED+** datasets. In the **"Positive"** category, emotions like **Joy (93%)**, **Surprise (75%)**, exhibit notably higher **positive correlations** when comparing **RAVDESS to AM-FED+**. This indicates a **stronger agreement** in the perception of these emotions as positive between the two datasets. Conversely, **Anger(53%)** and **Contempt (62%)** show **lower positive correlations**, suggesting

some variability in the interpretation of these emotions in the positive context between the two datasets.

2. **Likert Scale x Source**

   The Chi-square statistic of 211.34 with a p-value close to zero (1.2e-46) indicates a highly significant association between Likert scale categories and source datasets (RAVDESS or AM-FED). The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 12 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables.

3. **Source x Emotion**

   The Chi-square statistic of 0.0 with a p-value of 1.0 and 6 degrees of freedom indicates that there is no significant association between the Source and Emotion in the given contingency table.In that case we do not have enough evidence to reject the null hypothesis.

AM-FED+ vs Our-Collection

In consideration of the data presented in **Table 4.11 and 4.12** for AM-FED+ and Our-Collection, I conducted a comprehensive analysis to calculate the correlation between variables.

Interaction	Degree of freedom	p-value	chi-square	Reject or Accept
Likert Scale x Emotion	12	0	242.24	Rejected
Likert Scale x Source	2	0	33.53	Rejected
Source x Emotion	6	1	0	Not Rejected

Table 4.24: Correlation of Emotional Outputs for AM-FED+ vs Our-Collection Outputs

1. **Likert Scale x Emotion (AM-FED vs Our-Collection)** The Chi-square statistic of 242.24 with a p-value close to zero (5.6e-45) indicates a highly significant association between Likert scale categories and emotions. The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 12 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables. Certainly! The table represents the observed counts for each combination of the Likert Scale and Emotion. Each row corresponds to a Likert Scale category (Negative, Positive, or Zero), and each column corresponds to a specific emotion (Anger, Contempt, Disgust, Fear, Joy, Sadness, Surprise). The numbers in the table indicate how many data points fall into each combination. For instance, in the "Negative" Likert Scale category, there are 388 occurrences of Anger, 284 of Contempt, 180 of Disgust, and so on.This breakdown provides a detailed view of the distribution of emotions across different Likert Scale responses. The observed counts presented in the table make a total of 1059 videos considered, encompassing data from both the RAVDESS and AM-FED+ datasets.

Likert Scale	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise
Negative	120	175	123	116	18	80	86
Positive	335	279	336	349	494	392	382
Zero	91	92	87	81	34	74	78
Total	546	546	546	546	546	546	546

Table 4.25: Observed counts for each combination of Likert Scale and Emotion for AM-FED+ vs Our-Collection

Likert Scale	Anger	Contempt	Disgust	Fear	Joy	Sadness	Surprise
Negative	22%	32%	22%	21%	3%	15%	16%
Positive	61%	51%	62%	64%	90%	72%	70%
Zero	17%	17%	16%	15%	7%	13%	14%

Table 4.26: Percentage agreement between AM-FED+ vs Our-Collection

The table 4.26 illustrates the percentage agreement between the Likert Scale categories "Negative," "Positive," and "Zero" for each emotion when comparing the **AM-FED+** and **Our-Collection** datasets. In the **"Positive"** category, emotions like **Joy (90%)**, **Ssdness (72%)**, exhibit notably higher **positive correlations** when comparing **AM-FED+ to Our-Collection**. This indicates a **stronger agreement** in the perception of these emotions as positive between the two datasets. Conversely, **Contempt (51%)**, **anger (61%)** show **lower positive correlations**, suggesting some variability in the interpretation of these emotions in the positive context between the two datasets.

2. **Likert Scale x Source**

   The Chi-square statistic is a measure of the difference between the observed and expected frequencies in a contingency table. In this case, the Chi-square statistic is 33.53. The p-value associated with this statistic is 5.22e-08, which is much smaller than the conventional significance level of 0.05. The degree of freedom is 2, indicating the number of categories minus 1.Therefore, based on the p-value, we reject the null hypothesis of independence, indicating that there is a statistically significant association between the Likert Scale and Source categories in the observed data.

3. **Source x Emotion**

   The Chi-square statistic of 0.0 with a p-value of 1.0 and 6 degrees of freedom indicates that there is no significant association between the Source and Emotion in the given contingency table. In that case we do not have enough evidence to reject the null hypothesis.

Our-Collection vs Ravdess

In consideration of the data presented in **Table 4.12 and 4.10** for Our Collection and Ravdess, I conducted a comprehensive analysis to calculate the correlation between variables.

Interaction	Degree of freedom	p-value	chi-square	Reject or Accept
Likert Scale x Emotion	12	0	612.22	Rejected
Likert Scale x Source	2	0	48.94	Rejected
Source x Emotion	6	1	0	Not Rejected

Table 4.27: Correlation of Emotional Outputs for Our-Collection vs Ravdess Outputs

1. **Likert Scale x Emotion**

   The Chi-square statistic of 612.22 with a p-value close to zero (2.1e-123) indicates a highly significant association between Likert scale categories and emotions. The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 12 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables. Certainly! Table 4.28 represents the observed counts for each combination of the Likert Scale and Emotion. Each row corresponds to a Likert Scale category (Negative, Positive, or Zero), and each column corresponds to a specific emotion (Anger, Contempt, Disgust, Fear, Joy, Sadness, Surprise). The numbers in the table indicate how many data points fall into each combination. For instance, in the "Negative" Likert Scale category, there are 388 occurrences of Anger, 284 of Contempt, 180 of Disgust, and so on.This breakdown provides a detailed view of the distribution of emotions across different Likert Scale responses. The observed counts presented in the table make a total of 1059 videos considered, encompassing data from both the RAVDESS and AM-FED+ datasets.

Likert Scale	**Anger**	**Contempt**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**
Negative	318	151	97	82	12	148	72
Positive	344	498	567	581	687	510	598
Zero	47	60	45	46	10	51	39
Total	709	709	709	709	709	709	709

Table 4.28: Observed counts for each combination of Likert Scale and Emotion for Our-Collections vs Ravdess

Likert Scale	**Anger**	**Contempt**	**Disgust**	**Fear**	**Joy**	**Sadness**	**Surprise**
Negative	45%	21%	14%	11%	2%	21%	10%
Positive	48%	70%	80%	81%	97%	72%	84%
Zero	7%	9%	6%	8%	1%	7%	6%

Table 4.29: Percentage agreement between Our-Collection vs Ravdess

The table 4.29 illustrates the percentage agreement between the Likert Scale categories "Negative," "Positive," and "Zero" for each emotion when comparing the **RAVDESS** and **Our-Collection** datasets. In the **"Positive"** category, emotions like **Joy (97%)**, **Surprise (84%)**, exhibit notably higher **positive correlations** when comparing **RAVDESS to Our-Collection**. This indicates a **stronger agreement** in the perception of these emotions as positive between the two datasets. Conversely,

Anger(48%) show **lower positive correlations**, suggesting some variability in the interpretation of these emotions in the positive context between the two datasets.

2. **Likert Scale x Source**

   The Chi-square statistic of 211.34 with a p-value close to zero (2.3e-11) indicates a highly significant association between Likert scale categories and source datasets (RAVDESS and Our-Collection). The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 2 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables.

3. **Source x Emotion**

   The Chi-square statistic of 0.0 with a p-value of 1.0 and 6 degrees of freedom indicates that there is no significant association between the Source and Emotion in the given contingency table. In that case, we do not have enough evidence to reject the null hypothesis.

4.3.5 Activation Level Correlation

The earlier analysis in **section 4.3.4 Emotion Level Correlation** explored the relationships among dataset sources (RAVDESS, AM-FED+, Our-Collection), Likert scale categories (Negative, Positive, Zero), and various emotions using contingency tables and Chi-square tests. This examination provided insights into emotional expression in posed (RAVDESS), spontaneous (AM-FED+), and semi-spontaneous (Our-Collection) scenarios. The Chi-square test assessed significant associations, considering the likelihood of observed results by chance. The findings contribute to a nuanced understanding of emotion detection across diverse datasets and scenarios. Now, we extend this exploration to the **activation level**, applying the **same methodology** to gain insights into the correlation between **action units** in the specified datasets.

AM-FED+ vs Ravdess

Interaction	Degree of freedom	p-value	chi-square	Reject or Accept
Likert Scale x Action Unit	28	0	1217.81	Rejected
Likert Scale x Source	2	0	564.64	Rejected
Source x Action Unit	14	1	1.77	Not Rejected

Table 4.30: Correlation of Action Units Outputs for AM-FED+ vs Ravdess

1. **Likert Scale x Emotion**

   The Chi-square statistic of 1217.81 with a p-value close to zero (9.3e-239) indicates a highly significant association between Likert scale categories and emotions. The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 28 degrees of freedom, the analysis allows

for a comprehensive assessment of the interplay between these variables. Certainly! The table represents the observed counts for each combination of the Likert Scale and Emotion. Each row corresponds to a Likert Scale category (Negative, Positive, or Zero), and each row corresponds to a specific action unit . The numbers in the table indicate how many data points fall into each combination. For instance, in the "Negative" Likert Scale category, there are 141 occurrences of AU6, 41 of AU12, 5 of AU24, and so on. In the "Positive" Likert Scale category, there are 824 occurrences of AU6, 983 of AU12, and so on. This breakdown provides a detailed view of the distribution of action units across different Likert Scale responses. The observed counts presented in the table make a total of 1059 videos considered, encompassing data from both the RAVDESS and AM-FED+ datasets.

Action Unit	Negative	Positive	Zero	Total
AU1	201	767	91	1059
AU6	141	824	94	1059
AU4	227	730	102	1059
AU7	128	844	87	1059
AU9	92	823	93	1059
AU10	126	844	89	1059
AU12	41	983	35	1059
AU15	184	656	219	1059
AU17	109	807	101	1059
AU18	151	807	101	1059
AU20	408	514	137	1059
AU24	5	931	123	1059
AU26	100	885	74	1059
AU28	193	766	100	1059
AU43	147	798	114	1059

Table 4.31: Observed Counts for each combination of Likert Scale and Action units for Am-FED+ vs Ravdess

Percentage agreement between observed counts of Likert Scale and Action units for Am-FED+ vs Ravdess

Action Unit	Negative	Positive	Zero
AU1	19%	72%	9%
AU6	13%	78%	9%
AU4	21%	69%	10%
AU7	12%	80%	8%
AU9	9%	78%	13%
AU10	12%	80%	8%
AU12	4%	93%	3%
AU15	17%	62%	21%
AU17	10%	76%	14%
AU18	14%	76%	10%
AU20	39%	48%	13%
AU24	1%	88%	11%
AU26	9%	84%	7%
AU28	18%	72%	10%
AU43	14%	75%	11%

Table 4.32: Percentage agreement between observed counts of Likert Scale and Action units for Am-FED+ vs Ravdess

The table 4.32 illustrates the percentage agreement between the Likert Scale categories "Negative," "Positive," and "Zero" for each action units when comparing the **RAVDESS** and **AM-FED+** datasets. In the **"Positive"** category, action units like **AU12 (93%)**, **AU24 (88%)**, exhibit notably higher **positive correlations** when comparing **RAVDESS to AM-FED+**. This indicates a **stronger agreement** in the perception of these action units as positive between the two datasets. Conversely, **AU20(48%)** show **lower positive correlations**, suggesting some variability in the interpretation of these emotions in the positive context between the two datasets.

2. **Likert Scale x Source**

   The Chi-square statistic of 564.64 with a p-value close to zero (2.4e-123) indicates a highly significant association between Likert scale categories and source datasets (RAVDESS or AM-FED). The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 2 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables.

3. **Source x Emotion**

   The Chi-square statistic of 1.77 with a p-value of 1 and 14 degrees of freedom indicates that there is no significant association between the Source and Action Unit in the given contingency table. In that case, we do not have enough evidence to reject the null hypothesis.

AM-FED+ vs Our-Collection

Interaction	Degree of freedom	p-value	chi-square	Reject or Accept
Likert Scale x Action Unit	28	0	509.55	Rejected
Likert Scale x Source	2	0	133.55	Rejected
Source x Action Unit	14	1	0	Not Rejected

Table 4.33: Correlation of Action Units Outputs for AM-FED+ vs Our-Collection

1. **Likert Scale x Emotion**

   The Chi-square statistic of 509.55 with a p-value close to zero (7.2e-90) indicates a highly significant association between Likert scale categories and emotions. The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 28 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables. Certainly! The table represents the observed counts for each combination of the Likert Scale and Action Unit. Each row corresponds to a Likert Scale category (Negative, Positive, or Zero), and each row corresponds to a specific Action Unit. The numbers in the table indicate how many data points fall into each combination. For instance, in the "Negative" Likert Scale category, there are 141 occurrences of AU6, 41 of AU12, 5 of AU24, and so on. In the "Positive" Likert Scale category, there are 824 occurrences of AU6, 983 of AU12, and so on. This breakdown provides a detailed view of the distribution of action units across different Likert Scale responses. The observed counts presented in the table make a total of 546 videos considered, encompassing data from both the AM-FED+ and Our-Collection datasets.

Action Unit	Negative	Positive	Zero	Total
AU1	110	363	73	546
AU6	77	410	59	546
AU4	94	368	84	546
AU7	73	404	69	546
AU9	70	406	70	546
AU10	73	405	68	546
AU12	25	496	25	546
AU15	80	275	191	546
AU17	99	378	69	546
AU18	66	400	80	546
AU20	134	323	89	546
AU24	3	417	126	546
AU26	66	422	58	546
AU28	93	371	82	546
AU43	70	410	66	546

Table 4.34: Observed Counts for each combination of Likert Scale and Action units for Am-FED+ vs Our-Collection

Percentage agreement between observed counts of Likert Scale and Action units for Am-FED+ vs Our-Collection

Action Unit	Negative	Positive	Zero
AU1	20%	67%	13%
AU6	14%	75%	11%
AU4	17%	67%	16%
AU7	13%	74%	13%
AU9	13%	74%	13%
AU10	13%	74%	13%
AU12	5%	90%	5%
AU15	15%	50%	35%
AU17	18%	69%	13%
AU18	12%	73%	15%
AU20	24%	59%	17%
AU24	1%	76%	23%
AU26	12%	77%	11%
AU28	17%	68%	15%
AU43	13%	75%	12%

Table 4.35: Percentage agreement between observed counts of Likert Scale and Action units for Am-FED+ vs Our-Collection

Table 4.35 illustrates the percentage agreement between the Likert Scale categories "Negative," "Positive," and "Zero" for each action unit when comparing the **AM-FED+** and **Our-Collection** datasets. In the **"Positive"** category, action units like **AU12 (90%), and AU26 (77%)**, exhibit notably higher **positive correlations** when comparing **AM-FED+ to Our-Collection**. This indicates a **stronger agreement** in the perception of these action units as positive between the two datasets. Conversely, **AU20(59%)** shows **lower positive correlations**, suggesting some variability in the interpretation of these emotions in the positive context between the two datasets.

2. **Likert Scale x Source**

   The Chi-square statistic of 133.55 with a p-value close to zero (9.6e-30) indicates a highly significant association between Likert scale categories and source datasets (AM-FED+ and Our-Collection). The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 2 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables.

3. **Source x Emotion**

   The Chi-square statistic of 0 with a p-value of 1 and 14 degrees of freedom indicates that there is no significant association between the Source and Action Unit in the given contingency table. In that case, we do not have enough evidence to reject the null hypothesis.

Our-Collection vs Ravdess

In consideration of the data presented in **Tables 4.18 and 4.16** for Our Collection and Ravdess, I conducted a comprehensive analysis to calculate the correlation between variables.

Interaction	Degree of freedom	p-value	chi-square	Reject or Accept
Likert Scale x Action Unit	28	0	1204.36	Rejected
Likert Scale x Source	2	0	61.42	Rejected
Source x Action Unit	14	1	0	Not Rejected

Table 4.36: Correlation of Action Units Outputs for Our-Collections vs Ravdess Outputs

1. **Likert Scale x Emotion**

   The Chi-square statistic of 1263.32 with a p-value close to zero (1.96e-248) indicates a highly significant association between Likert scale categories and action units. The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 28 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables. Certainly! The table represents the observed counts for each combination of the Likert Scale and Action Unit. Each row corresponds to a Likert Scale category (Negative, Positive, or Zero), and each row corresponds to a specific Action Unit. The numbers in the table indicate how many data points fall into each combination. For instance, in the "Negative" Likert Scale category, there are 70 occurrences of AU6, 12 of AU12, 8 of AU24, and so on. In the "Positive" Likert Scale category, there are 598 occurrences of AU6, 681 of AU12, and so on. This breakdown provides a detailed view of the distribution of action units across different Likert Scale responses. The observed counts presented in the table make a total of 709 videos considered, encompassing data from both the AM-FED+ and Our-Collection datasets.

Action Unit	Negative	Positive	Zero	Total
AU1	101	570	38	709
AU6	70	598	41	709
AU4	149	520	40	709
AU7	63	614	32	709
AU9	85	591	33	709
AU10	75	601	33	709
AU12	16	681	12	709
AU15	128	537	44	709
AU17	30	635	44	709
AU18	99	563	47	709
AU20	330	293	86	709
AU24	8	694	7	709
AU26	50	629	40	709
AU28	114	545	50	709
AU43	83	570	56	709

Table 4.37: Observed Counts for each combination of Likert Scale and Action units for Ravdess vs Our-Collection

Percentage agreement between observed counts of Likert Scale and Action units for Our-Collection vs Ravdess

Action Unit	Negative	Positive	Zero
AU1	14%	80%	6%
AU6	10%	84%	6%
AU4	21%	73%	6%
AU7	9%	87%	4%
AU9	12%	83%	5%
AU10	10%	84%	6%
AU12	2%	96%	2%
AU15	18%	75%	7%
AU17	4%	90%	6%
AU18	14%	79%	7%
AU20	46%	41%	13%
AU24	1%	98%	1%
AU26	7%	89%	4%
AU28	16%	77%	7%
AU43	12%	80%	8%

Table 4.38: Percentage agreement between observed counts of Likert Scale and Action units for Ravdess vs Our-Collection

Table 4.38 illustrates the percentage agreement between the Likert Scale categories "Negative," "Positive," and "Zero" for each action unit when comparing the **Ravdess** and **Our-Collection** datasets. In the **"Positive"** category, action units like **AU24 (98%)**, **and AU12 (96%), AU17 (90%) and AU26(89%)**, exhibit notably higher

positive correlations when comparing **Ravdess to Our-Collection**. This indicates a **stronger agreement** in the perception of these action units as positive between the two datasets. Conversely, **AU20(41%)** shows **lower positive correlations**, suggesting some variability in the interpretation of these emotions in the positive context between the two datasets.

2. **Likert Scale x Source**

   The Chi-square statistic of 61.42 with a p-value close to zero (4.5e-14) indicates a highly significant association between Likert scale categories and source datasets (Ravdess and Our-Collection). The low p-value suggests strong evidence to reject the null hypothesis, implying that the observed relationships are unlikely due to chance. With 2 degrees of freedom, the analysis allows for a comprehensive assessment of the interplay between these variables.

3. **Source x Emotion**

   The Chi-square statistic of 0 with a p-value of 1 and 14 degrees of freedom indicates that there is no significant association between the Source and Action Unit in the given contingency table. In that case, we do not have enough evidence to reject the null hypothesis.

4.4 Conclusion

The existing literature appears to be in agreement regarding the intricate connection between action units and emotions. For example, author Dupre et al. (2019) affirm that anger expression evokes the engagement of AU4, AU7, AU10, and AU17) [39]. Nonetheless, contrary to this, according to their research, author S.velusamy (2011) depicts anger as a result of activation of AU2, AU4, AU7, AU17, and AU23. In contrast, author Lucey et al. (2010) argue that anger only comprises AU23 and AU24. [37].The same applies to other emotions as well. The problem is connecting action units and emotions as they develop from person to person. For example, though such anger AU4 and Anger AU7 correspond to anger, there is no full agreement reached concerning this. When there is no clear association of an emotion with a particular unit, other opinions are formed at different levels. For example, research in relation to AU4 and Anger links it with AU10 and AU17 among others. It's important to note that there's been a range of opinions among various studies regarding the connection between individual actions and emotions. This disagreement doesn't just happen when people are angry. It can happen with other emotions and the things that make them happen too. Investigating these subtleties helps in developing a deep understanding of how facial expressions and other complexities interact among emotions. We have shown the relationship between Dominant Action Units for each Emotion in chapter 4.

5 Conclusions

This research endeavored to examine the evaluations of emotional responses by facial emotion recognition systems, and to examine their effect on various variables, such as the nationality, age, and gender of the speaker. We started by examining the outcomes of the performance comparison between the posed, spontaneous, and semi-spontaneous models. The results are also statistically significant, showing that there are differences between the two systems (Emotient and Affectiva), which could be due to the performance of the two systems.

1. The two systems, namely Emotient and Affectiva, are not universally applicable as certain emotions, such as joy, are accurately detected while others are not. Furthermore, some action units are detected correctly, such as AU12, and others are not.
2. It is possible that the issue could be attributed to the architecture of Emotient and Affectiva. The pipeline for both systems is different because each system uses different algorithms to process face-recognition videos.
3. The training database for both systems was different, as was the racial mix, gender, and age. The training data for one system is recorded in a laboratory setting, while the other system is collected by crowdsourcing, with the emotions recorded over the internet. These are two different datasets in terms of people and technology.
4. The conflict between the systems that is demonstrated at the Action unit level can be explained by two systems differing in the mapping from one action unit emotion to another. Either the emotion is made up of a certain linear combination of Action Units or it may be a combination of both. The way other features are differentiated between the two systems is related to these differences.

6 Future Works

There are three different ways to recognize emotions: video, audio, and text. We have done video (Facial Emotion Recognition) and we can expand it to audio and texts. We can combine all three modalities. Open smile and open Vokaturi are two ways to recognize emotions in speech. Many people use these systems to recognize emotions. Open Vokaturi had been trained on SAVEE, but Open Smile had not been. Second, it would be fascinating to contrast the Posed and Spontaneous datasets, which could provide a clearer picture of how diverse populations express their emotions. It would be highly advantageous to establish a method or algorithm for integrating speech and head movements, including yaw, pitch, and roll degrees, and devise a method or algorithm for combining all three modalities, namely face, speech, and text.

Bibliography

[1] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 298–305.

[2] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[3] T. Akter, M. H. Ali, M. I. Khan, M. S. Satu, M. J. Uddin, S. A. Alyami, S. Ali, A. Azad, and M. A. Moni, "Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage," *Brain Sciences*, vol. 11, no. 6, p. 734, 2021.

[4] C. Vogel and K. Ahmad, "Agreement and disagreement between major emotion recognition," *Available at SSRN 4431888*, 2023.

[5] D. McDuff, R. El Kaliouby, and R. W. Picard, "Crowdsourcing facial responses to online videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 456–468, 2012.

[6] 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, p. e0196391, 2018.

[7] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the national academy of sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[8] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:3329621

[9] S. Velusamy, H. Kannan, B. Anand, A. Sharma, and B. Navathe, "A method to infer emotions from facial action units," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 2028–2031.

[10] J. P. Mitchell, "Inferences about mental states," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1309–1316, 2009.

[11] R. El Kaliouby and P. Robinson, "Mind reading machines: Automated inference of cognitive mental states from video," in *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, vol. 1. IEEE, 2004, pp. 682–688.

[12] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[13] D. McDuff, J. M. Girard, and R. e. Kaliouby, "Large-scale observational evidence of cross-cultural differences in facial behavior," *Journal of Nonverbal Behavior*, vol. 41, pp. 1–19, 2017.

[14] D. McDuff, M. Amr, and R. El Kaliouby, "Am-fed+: An extended dataset of naturalistic facial expressions collected in everyday settings," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 7–17, 2018.

[15] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 568–573.

[16] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[17] R. W. Picard, "Measuring affect in the wild," in *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*. Springer, 2011, pp. 3–3.

[18] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, 2016, pp. 3723–3726.

[19] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, J. R. Movellan *et al.*, "Automatic recognition of facial actions in spontaneous expressions." *J. Multim.*, vol. 1, no. 6, pp. 22–35, 2006.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. Ieee, 2001, pp. I–I.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[22] J. M. Garcia-Garcia, V. M. Penichet, and M. D. Lozano, "Emotion detection: a technology review," in *Proceedings of the XVIII international conference on human computer interaction*, 2017, pp. 1–8.

[23] W. V. Friesen, P. Ekman *et al.*, "Emfacs-7: Emotional facial action coding system," *Unpublished manuscript, University of California at San Francisco*, vol. 2, no. 36, p. 1, 1983.

[24] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in health sciences education*, vol. 15, pp. 625–632, 2010.

[25] J. Murray, "Likert data: what to use, parametric or non-parametric?" *International Journal of Business and Social Science*, vol. 4, no. 11, 2013.

[26] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi medical journal*, vol. 24, no. 3, pp. 69–71, 2012.

[27] C. P. Dancey and J. Reidy, *Statistics without maths for psychology*. Pearson education, 2007.

[28] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010.

[29] E. Ostertagova, O. Ostertag, and J. Kováč, "Methodology and application of the kruskal-wallis test," *Applied mechanics and materials*, vol. 611, pp. 115–120, 2014.

[30] R. F. Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.

[31] D. W. Zimmerman and B. D. Zumbo, "Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks," *The Journal of Experimental Education*, vol. 62, no. 1, pp. 75–86, 1993.

[32] H.-Y. Kim, "Statistical notes for clinical researchers: Chi-squared test and fisher's exact test," *Restorative dentistry & endodontics*, vol. 42, no. 2, pp. 152–155, 2017.

[33] B. S. Everitt, *The analysis of contingency tables*. CRC Press, 1992.

[34] D. Sharpe, "Chi-square test is statistically significant: Now what?" *Practical Assessment, Research, and Evaluation*, vol. 20, no. 1, p. 8, 2015.

[35] K. A. Batterton and K. N. Hale, "The likert scale what it is and how to use it," *Phalanx*, vol. 50, no. 2, pp. 32–39, 2017.

[36] S. Du and A. M. Martinez, "Compound facial expressions of emotion: from basic research to clinical applications," *Dialogues in clinical neuroscience*, 2022.

[37] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[38] M. L. McHugh, "The chi-square test of independence," *Biochemia medica*, vol. 23, no. 2, pp. 143–149, 2013.

[39] D. Dupré, E. Krumhuber, D. Küster, and G. McKeown, 2019.