Estimating user's emotion from non-verbal behaviour in communication

Yuichi Midorikawa

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

A dissertation presented in part fulfilment of the requirements of the Degree of Master of Science at The University of Glasgow

December 13, 2021

Abstract

Interactions between humans and computers will become more natural as computers perceive non-verbal communication, such as human emotions, and modify their conversational strategies. This paper focuses on developing a model for estimating users' emotions in communication from multi-modal features to support application users with intelligent agents. For this purpose, we use a multimodal dataset containing features of the face, eyes, mouth and posture of participants observed in presentations and conversations in 14 videos. The dataset also contains English subtitles of what each participant said, and we perform sentiment analysis using a pre-trained BERT model to prepare user emotion categories. We extracted various features such as face, eyes, mouth and posture movements to estimate the emotional category of the user. From these features, we created classification models for inferring Ekman-level and Group-level emotional categories using Pytorch's three-layer neural network and random forest and evaluated the accuracy of the emotional category estimation. As a result of our experiments, multimodal models achieved a classification accuracy of 0.82 at the Ekman level and 0.96 at the Group level. Despite the limited amount of non-verbal information in the dataset used in this project, we believe that the model has an excellent potential to predict users' emotions. In addition, we were able to collect non-verbal data from a video of a user talking to create an emotion estimation model. Using the same method for other videos, we obtained non-verbal information about the user's face, hands, mouth, and posture, which we can use in future projects.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: ______________________ Signature: ________________________

Acknowledgements

I would like to express my sincere gratitude to Professor Nicolas Pugeault for being helpful and supportive throughout this project. The advice, knowledge, and guidance he shared with me were vital to completing this project.

I would also like to thank my family and friends who have been a great support throughout this year.

Contents

Chapter 1

Introduction

1.1 Context

The findings of conversation analysis in sociology and communication science have revealed the importance of the verbal information exchanged during face-to-face conversations and the role of non-verbal information such as speech, gestures, posture, gaze, and facial expressions [1]. For example, communication between people or between people and virtual agents involves spoken words and non-verbal information such as hand gestures, facial expressions, and tone of voice to express emotions and provide feedback. For example, the virtual human agent Rachel can exchange non-verbal communication through facial expressions and body gestures [2]. In addition, engagement is an essential part of the communication process. Engagement is the process by which a perceived connection between two or more participants in a conversational interaction is established, maintained and terminated [3]. For example, to maintain a conversation, a speaker will ensure that the listener is paying attention and participating appropriately in the discussion. On the other hand, the listener uses non-verbal behaviours such as gaze and nodding to show attention to the conversation and convey that they are willing to maintain communication.

Figure 1.1: Character conversation AI technology [2]

Therefore, while both sides need to be aware of each other's engagement in face-to-face conversations, the engagement must also be considered when designing interactions between people and computers, such as virtual agents. It is no exaggeration to say that an interface that ignores the human emotional state in interaction and does not react appropriately to that state will never be able to gain trust. If computers could perceive and respond to non-verbal communication, such as human emotions, the interaction between humans and computers would be more natural. It is also necessary to estimate emotions from the non-verbal information of the user to establish engagement to achieve natural interaction between the user and the computer. Therefore, it is worth evaluating a speaker's emotions from non-verbal information such as the face, mouth, eye and posture.

In addition to this background, with advances in computer vision and human sensing technology, there has been much research into the automatic detection and recognition of non-verbal information. Several automated emotion recognition systems use facial expressions [4], or posture [5] to detect human emotional states, but relatively few have focused on emotion recognition using both modalities. A multimodal approach should result in better performance and greater robustness when one of these modalities comes into play in a noisy environment.

1.2 Overall Objectives

This project will support application users with intelligent agents by estimating and assessing users' emotions from non-verbal information such as the externally observable facial mouth and eye landmarks and posture in communication. This project will look to create models to estimate the following emotions. Compare these models using the same data set and metrics.

1. A model predicts emotion categories per frame from features related to face, mouth and eye landmarks.
2. A model predicts emotion categories per frame from features related to body posture.
3. A model predicts emotion categories per frame from features related to the face, mouth, eye landmarks, and body posture.

1.3 Report Structure

This paper consists of five chapters. Chapter 1 gives a brief introduction to the necessity of engagement in interaction design and the objectives of this project. Chapter 2 describes the background and issues to be addressed in more detail and explains the related studies and techniques. Chapter 3 provides details of the design and implementation, including some of the challenges faced in the process. Chapter 4 discusses the details of testing and evaluation. Chapter 5 is the conclusion, including results, reflections and directions for future work.

Chapter 2

Analysis and Requirements

2.1 Problem Statement

In research on Human-Computer Interaction, one of the critical challenges is to achieve a smoother exchange of information between people and computers. In particular, interactive systems, which aim to communicate with verbal communication, are expected to be highly usable interfaces that allow users to interact with the systems in natural language. More recently, Smartphones have become increasingly popular with voice-recognition applications that will enable users to search and control information by voice. Spoken dialogue systems, or spoken language dialogue systems, are expected to find a broader range of applications in the coming years. On the other hand, the dramatic advances in computer graphics have improved the expression of animated characters. They have driven research into interface agents that can mediate human and computer interaction. The interface agents have a physical appearance, using their faces and bodies to communicate and show non-verbal behaviours. Therefore, spoken dialogue systems can function as a speech interface on their own and be a component technology of humanoid interfaces by being embedded as a voice dialogue module in animated conversational agents or communication robots. In such humanoid interfaces, the communication modalities are even more diverse. The system is an anthropomorphic entity and can use spoken words and non-verbal information such as hand gestures, facial expressions, and tone of voice. It is vital for interactive intelligent systems in the future that virtual agents have a human-like appearance and interface.

Recent research on virtual agents and communicative robots has shown that engagement through conversation is fundamental and essential for communication between users and humanoid interfaces [6]. As defined in [3], engagement is the process by which two or more participants can establish, maintain, and terminate a recognised connection. Engagement is a combination of verbal communication and non-verbal behaviour, supporting the bond between participants in a conversation. While verbal communication provides detailed and rich semantic information and social connections, non-verbal communication includes information on what has been understood so far, on the interests of each (or together) interactant, on the weakening of links and on the intention to disengage. For example, it is no exaggeration to say that an interface that ignores the human emotional state of interaction and does not react appropriately to that state will never be trusted. Instead, the user will regard the virtual agent as ”cold”, ”untrustworthy”, and ”unsocial”. If the user is no longer fully engaged in the conversation, then the information presented by the virtual agent is not adequately communicated to the user. Suppose computers can recognise the emotional input of the user. In that case, they will provide specific and appropriate assistance in communication according to the user's needs and preferences. Therefore, to establish a natural interaction between the user and the virtual agent, it is essential to estimate emotions from non-verbal information such as the user's facial expression and posture to show that the virtual agent is listening to the user. This task of inferring a speaker's emotions from non-verbal information such as facial expressions and posture is a practical way to establish engagement in communication.

With advances in computer vision and human sensing technology, there has been much research into the automatic detection and recognition of non-verbal information. Consequently, recognising facial expressions and body posture is a rapidly growing sub-field of image processing. Some studies have started to estimate higher-level communicative characteristics in recent years by integrating single nonverbal information such as gaze and gesture as multimodal communication [7]. The importance of incorporating multiple modalities in dialogue systems has shown itself in studies of multimodal communication understanding and response generation [8]. The multimodal approach expects to be more accurate and more robust when only one of the modalities is acquired in a noisy environment [9]. With this background, this project will focus on estimating and evaluating users' emotions from externally observable non-verbal information such as face and body posture in communication to support application users with intelligent agents.

2.2 Related Work

2.2.1 Emotion Classification Models

Psychology and neuroscience have extensively researched how we express and perceive emotions. One significant result is that the perception of emotions is at least categorised to some extent [10] [11]. The perception of tagged emotions assumes that the perceptual boundaries between categories change sharply rather than slowly. The datasets used in many studies of emotion recognition include the six basic emotion categories proposed by Ekman [12] (joy, anger, fear, sadness, disgust, and surprise) and the circumplex model of emotion that underpins emotional states [13]. The six categories proposed by Ekman are said to be universal. The theory of emotion categories is that different cultures use identical facial muscles to produce emotions. In addition, studies using carefully designed human observers have established types of emotion recognition [14].

We often use feature-based or neural network models to classify emotions automatically. Feature-based models often use hand-built dictionaries to annotate more details, such as the Valence Arousal Dominance Lexicon [15]. Pre-trained transformer-based models of language models, using representations of the neural network model BERT [16], have recently proven to reach state-of-the-art performance on several NLP tasks, including sentiment prediction. The models of the top performers in the Emotion Challenge [17] are all pre-trained BERT models.

We referred to the types of emotion taxonomy defined from a review of GoEmotion research [18]. Demszky et al. annotated 27 emotions and neutrals based on Semantic Space Theory using GoEmotions, a dataset of over 58,000 English Reddit comments, to train NLP models in an emotion recognition task. They achieved an average F1 score of 0.46 on the 28 emotion classification tasks, fine-tuning the pre-trained BERT model. In addition, they grouped emotions hierarchically and evaluated the model's performance at each hierarchy level. For example, there are four labels at the sentiment level (group level): positive, negative, ambiguous and neutral. Another taxonomy at the Ekman level uses the Neutral label to further divide the taxonomy into the following groups: Anger, Disgust, Fear, Joy, Sadness and Surprise. Each F1 score achieved 0.69 for the Sentiment level and 0.64 for the Ekman level. Therefore, This project will use the same BERT-based model pre-trained and classified at the Sentiment and Ekman levels.

2.2.2 Face Recognition

Researchers have been working on face recognition since the 1970s [19], and face recognition systems use face detection to separate faces given an input image of multiple faces. The pre-process each face to obtain a low-dimensional representation. Low-dimensional representations are essential for efficient classification. Face recognition is a challenge because images may capture different perspectives of a face. Facial expressions need to distinguish between interpersonal image changes between other people while being robust to intra-individual image changes such as age, presentation and style [20].

Jafri and Arabnia [21] provide a comprehensive survey of face recognition techniques until 2009. According to this survey, there are two approaches to face recognition research: a feature-based approach and a holistic approach. The early work in face recognition was feature-based and attempted to explicitly define low-dimensional face representations based on distance, area and angle [19]. Explicitly defined face representations are desirable for intuitive feature spaces and methods. However, explicitly defined terms have not been accurate. Therefore, they searched for a holistic approach, using statistics and artificial intelligence to learn from and achieve good results on face image datasets.

Statistical methods such as Principal Component Analysis (PCA) [22] represent faces as combinations of eigenvectors [23]. Face recognition methods based on PCA include Eigenfaces [24] and Fisherfaces [25]. Lawrence et al. [26] introduce an AI technique that uses convolutional neural networks to classify images of faces. The face recognition technology using convolutional neural networks is still one of the best in the world. Therefore, face recognition technology is mainly based on convolutional neural networks. According to [27], OpenFace has obtained the accuracy and performance of state-of-the-art deep learning techniques. This open-source tool provides facial landmark detection, head pose estimation, facial action unit recognition and gaze estimation. It uses a simple 2D matrix to transform the coordinates so that eyes and nose appear similar in the neural network input. This project will use this OpenFace to obtain features related to the face, mouth, and eyes landmarks.

2.2.3 Emotion Recognition by Facial Features

Facial expressions are an essential cue for emotions. For this reason, there are several approaches to classifying human emotional states. The features used in the classification task relate to a specific spatial position or displacement of a particular point or region of the face. This study [28] built an emotion recognition system using the meaningful orientation of specific facial muscles. They manually placed 11 windows on the face and used optical flow to extract the muscle movements. For classification, they used the K-nearest neighbour rule and obtained an accuracy of 80% for four emotions: joy, anger, sadness and anger. Instead of using facial muscle movements, Yacoob et al. [4] built a dictionary to convert activities associated with the edges of the mouth, eyes and eyebrows into a frame-by-frame mid-level linguistic representation. They used a rule-based system to classify the six basic emotions (happiness, surprise, anger, disgust, fear and sadness) with 88% accuracy for a sample size of 105. Black et al. used parametric models to extract the shape and movement of the mouse and eye [29]. They have constructed medium to high-level representations of facial actions and obtained an accuracy of around 78.5% for six emotions (happiness, sadness, anger, fear, surprise and disgust) across 36 video flips, including talk shows, news and movies.

Despite the remarkable success of traditional face recognition methods by extracting hand-crafted features, researchers have turned towards deep learning approaches for their high automatic recognition capabilities for the past decade. In 2019, Agrawal et Mittal [30] studied the impact of varying CNN parameters in recognition rate with the FER2013 database. First, all images were 64x64 pixels and varied the size and quantity of filters and the optimiser type chosen (SGD) for a simple CNN containing two successive convolutional layers. The second layer acted as the maximum pooling and utilised a softmax function to classify the input values. These studies have created two new CNN models and achieved an average accuracy of nearly 65% in recognising seven emotions (anger, disgust, fear, joy, sadness, surprise, and neutral).

2.2.4 Pose Detection

Classical methods for estimating human multi-joint posture combine observations of body parts with spatial relationships between those parts to make inferences. The spatial model of the articulated pose offers a graphical model of the tree structure [31] on which to base it. This method is successful for images where all the person's limbs are visible. Still, it is prone to characteristic errors such as double counting of image evidence caused by correlations between variables that the tree structure model does not cover. There is also a model based on non-tree structures [32], which adds edges such as symmetry, occlusion and long-range relations to the tree structure. These methods have to make a trade-off between an accurate model of the spatial relationships and a model that allows efficient hypothesis, often using a simple parametric form that allows fast inference. Convolutional neural networks (CNNs) have received widespread attention and greatly improved body pose estimation accuracy [33]. Therefore, in this project, this project will use OpenPose with CNNs.

2.2.5 Emotion Recognition by Body Pose

Facial expressions and body movement patterns complement each other in visual perception [34]. The movements of our bodies reflect our inner emotional and affective states [35]. For example, when a person is angry, they may walk fast. In recent years, progress in computer vision and machine learning has made it possible to efficiently recognise emotions from faces, voices, objects, and images. However, studies on emotional recognition from body movements have received less attention. One possible reason for this is conflicting opinions about how body movement patterns transmit emotional information.

Physical expressions, such as posture and movement patterns, play an essential role in understanding emotions. Coulson's study revealed the effect of body posture on the transmission of emotion by semantic analysis of static images [36]. In the experiment, observers categorised the body posture images into emotions and associated sadness with a tilted head and anger with a bent elbow. Thus, this shows that the head, chest, elbows, and shoulders movements represent essential features of physical emotions. Body movement models allow us to think of whole-body actions requiring gestures, posture changes, and weight distribution. In recognising emotions from body patterns, the kinematics of movement (such as velocity and acceleration) provide an essential feature [37].

Kleinsmith et al. [5] reported significant results when recognising emotions from body posture. They tested a system based on the CALM (categorising and learning module) [38] with 212 poses in 9 emotional states (Angry, Confused, Fear, Happy, Interest, Relaxed, Sad, Startled, Surprised). The experiment data showed that the overall accuracy of the dynamic postures was 70%. The CALM network consists of several modules and incorporates a neural network like the brain. Schindler et al. [39] introduced an image-based recognition system for emotions from images of body postures. His system's overall recognition accuracy was 80% for the seven basic feelings (angry, disgusted, fearful, happy, neutral, sad and surprised). This project will also focus on postural features and emotion recognition.

2.3 Related Technologies

2.3.1 BERT

The various natural language processing tasks involved come in two broad categories: token-level and sentence-level. Whereas token-level tasks involve learning relationships between tokens, sentence-level tasks involve learning connections between sentences. For both functions, pre-learning is helpful, and Pre-training is learning useful features for application to various tasks. There are two approaches to pre-learning: feature-based and fine-tuning approaches. In the former case, N-gram models and Word2Vec, typical in natural language processing, apply the features obtained by pre-training to various NLP tasks. The latter approach involves pre-training a language model to predict the next word, followed by task-specific supervised learning to fine-tune it. BERT models are pre-training models with fine-tuning of the latter, specifically using a bi-directional transformer.

The transformer is a neural translation model that uses attention and performs better than LSTM with little training. Attention is a technique for focusing attention on important parts of a series of data and includes additive attention, productive attention and self-attention. Self-Attention is generic and effective and performs well in all language processing tasks [40]. BERT model uses several units of this transformer.

One feature of the BERT model is learning about the context before and after the word of interest (bidirectional). Whereas previous models could only use the previous context (unidirectional) due to the problem of word anticipation, the BERT model can learn in both directions by devising a pre-learning task. This project will also use the pre-trained BERT model to analyse and extract sentiment categories.

2.3.2 OpenFace

This project uses OpenFace [41] to estimate the 3D position and gaze direction from 2D video information from Youtube videos. OpenFace follows the Facial Action Coding System (FACS) of Ekman et al. to obtain the head and gaze directions, the Action Units (AUs) characterising each part of the face, and the confidence level of the estimation results. OpenFace outputs two types of AU: a quantity representing the intensity of each AU in six steps (AUr) and a binary value representing whether or not an AU is present (AUc). OpenFace also outputs 17 different types of AU for AUr and 18 different types of AU for AUc. By Applying OpenFace to each frame of the face image in each trial, we can obtain a time series of AU changes.

2.3.3 OpenPose

This project will use OpenPose [42], which can estimate the 2D pose of a person from 2D video information from Youtube videos. OpenPose is a pose estimation library developed by the University of California, Berkeley and Carnegie Mellon University that uses deep learning to extract critical information about human joints and other features in real-time. The system can detect a person's body in a video or image and 135 crucial points of the face and hands. The part of this system is that it can be analysed using only pictures and videos from the camera, without using special sensors such as accelerometers. GPUs also allow a fast analysis of images and videos with multiple people in them.

Chapter 3

Design and Implementation

3.1 Overview (Research Steps, Development Environment)

This chapter builds on the theoretical account given in the previous chapters. It describes in detail the process from data collection to the implementation of the model to predict emotion categories. Firstly, we will describe the research steps and the execution environment to visualise the whole research process.

The following section then follows the sequence of the research steps, firstly describing the data collection and pre-processing in this project. For example, we will focus on performing sentiment analysis using a pre-trained BERT model on data classified by sentiment categories at the Group and Ekman levels and feature extraction using OpenFace and OpenPose, which are critical technologies in this project. We will also focus on feature extraction using OpenFace and OpenPose, key technologies in this project. This section also explains how to implement the model for estimating emotional categories. Furthermore, we will discuss the evaluation metrics and results in the next chapter.

3.1.1 Research Steps

We developed the following research steps to realise this project, aiming to estimate and evaluate users' emotions from externally observable non-verbal information such as facial expressions and postures in communication.

This research step consists of three phases. Phase 1, Data Preparation, prepares the training and evaluation data from the raw data and supports the next and subsequent phases (Phase 2, Implementation and Phase 3, Validation) to train and evaluate the model. Raw data uses YouTube videos and English subtitles to obtain data that includes people talking and text.

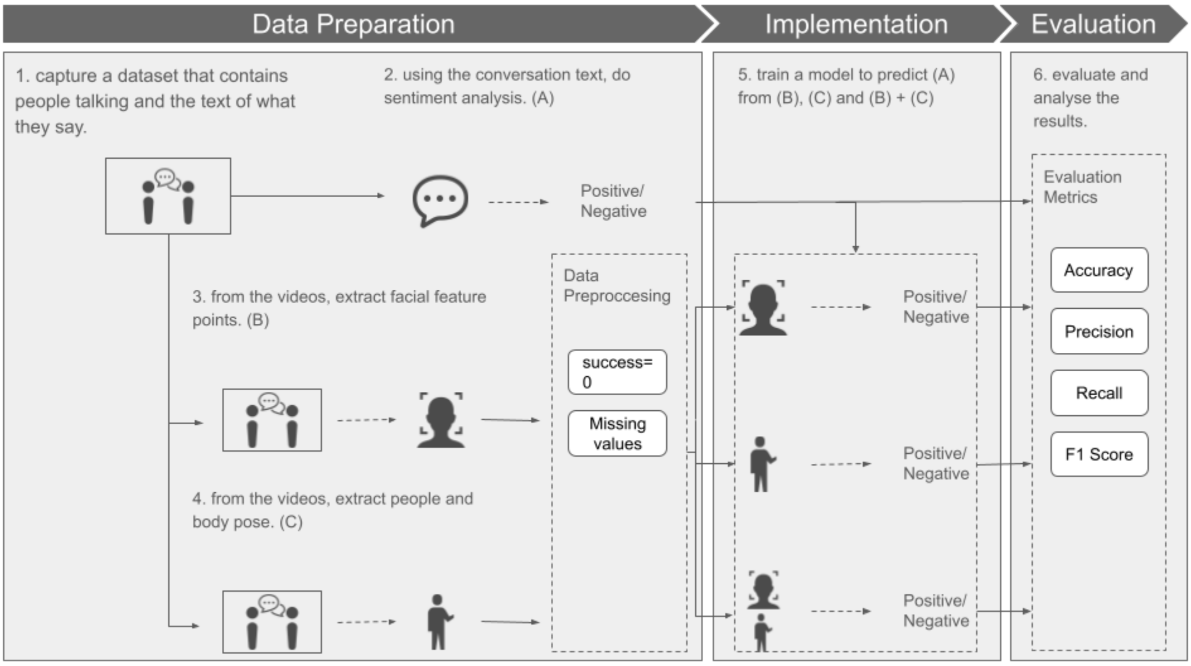

Figure 3.1: Research steps in the project

The selection criteria for the YouTube videos mentioned above are as follows.

- While the speaker is speaking, their facial expression and posture are observable.
- They speak with as much emotional expression as possible, which the author could judge subjectively.
- English subtitles available.

To prepare the sentiment categories to be used as the objective function for the training and evaluation data, we conducted a sentiment analysis on each sentence of the English subtitles of the YouTube video. We extracted the sentiment categories (referring to (A) of the research steps). We used a pre-trained BERT-based model with emotional categories classified at the Ekman and Group levels for its sentiment analysis.

Next, to prepare the facial expression and posture features as explanatory functions for the training and evaluation data (referring to step (B) of the research), we used FFmpeg on YouTube videos to split the subtitles into single sentences. For the segmented video, we used OpenFace and OpenPose to extract features related to face, eye, mouth and posture. In addition, the features used as training and evaluation data underwent data pre-processing, as the observed data contained missing values and frames with incorrect feature extraction.

In Phase 2, Implementation, we create models to estimate the user's emotions from externally observable non-verbal information such as facial landmarks and posture. The model built in this project is as follows (B) and (C) below correspond to the research steps) .

- (B) A model for predicting emotion categories per frame from features related to landmark expressions of the face, mouth and eyes extracted using OpenFace.
- (C) A model for predicting emotion categories per frame from features related to posture extracted using OpenPose.
- (B) + (C) a model for predicting emotion categories per frame from features extracted using OpenFace and OpenPose.

We used of a variety of open-source machine learning libraries to create our models. We use these models to validate and compare the accuracy of our sentiment category predictions.

In Phase 3 Evaluation, we compared and evaluated these models.

3.1.2 Development Environment

We used the Colab Pro version of Google Colaboratory (Colab) for this project because we needed a GPU to process OpenPose and build neural network models using Pytorch. Colab is a service provided by Google Research that allows anyone to write and run Python in a browser, making it particularly suitable for machine learning, data analysis and education. Colab Pro gives us priority access to Google's GPU and memory (25.46GB).

3.2 Data Preparation

3.2.1 Data Collection

This project collected fourteen videos with an average length of 13 minutes and 59 seconds. The average number of utterances by the speaker was 269 per video, for a total of 3761. In other words, we were able to obtain 3761 emotional categories. The list of fourteen videos describes appendix A. We built a YouTube Downloader with Python, using the Towards Data Science website as a reference [43].

To get English subtitles for a YouTube video, play the video that needs English subtitles and click on the three horizontal dots at the bottom of the video. Then select ”Open transcript”, and the transcript will appear on the right side of the video. The English subtitles and the timestamp (start time) for each sentence can be retrieved there.

From the timestamp of the English subtitle, determine the end time of each sentence and the interval (end time - start time) during that sentence. In order to perform OpenFace and OpenPose feature extraction on the video for each sentence, cut out the YouTube video for each sentence using FFmpeg, which is a free software for recording, converting, and playing back video and audio. Using the timestamp for each sentence and the interval during the speech, extract a video for each sentence and output it in (a) mp4 for OpenFace input or manual data observation, and (b) AVI file format for OpenPose input.

3.2.2 Conducting Sentiment Analysis

In order to prepare sentiment categories to be used as objective functions for training and evaluation data in this project, conducted sentiment analysis on a single sentence of English subtitles. We group the sentiments hierarchically and evaluate the model's performance at each hierarchy level. The emotional categories for this project referred to work defined from a review of GoEmotion research, based on Semantic Space Theory, annotated for 27 emotion and neutral types. The authors fine-tuned the BERT language model to achieve an average F1 score of 0.46. They achieved an F1-score (macro-average) of 0.64 for Ekman-level grouping into six coarse categories with their taxonomy by fine-tuning the BERT-based model and 0.69 for Group-level emotion grouping. This project will also use the same BERT model that has been classified and trained on its Group (Hierarchical Grouping) and Ekman levels using pipeline [44].

Table 3.1: Results by emotion taxonomy [18]

Taxonomies	Emotional categories	F-Score
Hierarchical Grouping	positive, negative, ambiguous + neutral	0.69
Ekman	anger, disgust, fear, joy, sadness, surprise + neutral	0.64
Original GoEmotions	27 emotions + neutral	0.46

An example output is shown below, showing that it is possible to get one or more sentiment categories and their confidence levels for a single sentence. Here only the output for the input is described. For the text "I love you, brother.", a sentiment analysis would output an emotion category of joy and a confidence value of 0.9991.

Listing 3.1: a Sample Result of Sentiment Analysis [44]

```
# This is an input text
texts = [
    "I love you, brother.",
]
# This is an output
[{'labels': ['joy'], 'scores': [0.99910116]}]
```

The results of the sentiment analysis for the 14 videos are as follows. There is a bias in the sentiment categories. There is a bias in the sentiment categories. The results show that neutral accounts for 65.59% of the Ekman level and 69.95% of the Group level.

Video	Sentence	neutral		joy		surprise		anger		sadness		fear		disgust	
Name	Number	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%	Number	%
video01	209	129	61.72%	52	24.88%	17	8.13%	6	2.87%	4	1.91%	1	0.48%	0	0.00%
video02	228	145	63.60%	41	17.98%	16	7.02%	17	7.46%	7	3.07%	0	0.00%	2	0.88%
video03	405	288	71.11%	53	13.09%	18	4.44%	32	7.90%	12	2.96%	2	0.49%	0	0.00%
video04	212	148	69.81%	32	15.09%	18	8.49%	1	0.47%	12	5.66%	1	0.47%	0	0.00%
video05	284	178	62.68%	71	25.00%	20	7.04%	6	2.11%	5	1.76%	4	1.41%	0	0.00%
video06	327	185	56.57%	62	18.96%	66	20.18%	6	1.83%	6	1.83%	2	0.61%	0	0.00%
video07	258	157	60.85%	80	31.01%	13	5.04%	3	1.16%	4	1.55%	1	0.39%	0	0.00%
video08	338	231	68.34%	66	19.53%	28	8.28%	8	2.37%	2	0.59%	2	0.59%	1	0.30%
video09	292	172	58.90%	45	15.41%	18	6.16%	42	14.38%	10	3.42%	3	1.03%	2	0.68%
video10	291	192	65.98%	46	15.81%	18	6.19%	17	5.84%	11	3.78%	5	1.72%	2	0.69%
video11	192	119	61.98%	53	27.60%	8	4.17%	8	4.17%	3	1.56%	1	0.52%	0	0.00%
video12	252	172	68.25%	52	20.63%	13	5.16%	9	3.57%	5	1.98%	1	0.40%	0	0.00%
video13	306	227	74.18%	41	13.40%	18	5.88%	11	3.59%	7	2.29%	0	0.00%	2	0.65%
video14	167	124	74.25%	26	15.57%	7	4.19%	0	0.00%	1	0.60%	9	5.39%	0	0.00%
Total	3761	2467	65.59%	858	22.81%	278	7.39%	215	5.72%	112	2.98%	43	1.14%	9	0.24%

Figure 3.2: Distribution of Ekman-level emotional categories

Video	Sentence	neutral		positive		ambiguous		negative	
Name	Number	Number	%	Number	%	Number	%	Number	%
video01	209	133	63.64%	48	22.97%	15	7.18%	13	6.22%
video02	227	153	67.40%	36	15.86%	17	7.49%	21	9.25%
video03	403	303	75.19%	42	10.42%	16	3.97%	42	10.42%
video04	212	163	76.89%	24	11.32%	12	5.66%	13	6.13%
video05	283	201	71.02%	47	16.61%	17	6.01%	18	6.36%
video06	327	193	59.02%	58	17.74%	61	18.65%	15	4.59%
video07	258	170	65.89%	65	25.19%	13	5.04%	10	3.88%
video08	338	243	71.89%	49	14.50%	31	9.17%	15	4.44%
video09	292	194	66.44%	23	7.88%	18	6.16%	57	19.52%
video10	291	201	69.07%	38	13.06%	18	6.19%	34	11.68%
video11	192	127	66.15%	51	26.56%	5	2.60%	9	4.69%
video12	252	181	71.83%	44	17.46%	12	4.76%	15	5.95%
video13	306	232	75.82%	33	10.78%	18	5.88%	23	7.52%
video14	167	134	80.24%	16	9.58%	5	2.99%	12	7.19%
Total	3757	2628	69.95%	574	15.28%	258	6.87%	297	7.91%

Figure 3.3: Distribution of Group-level emotional categories

3.2.3 Extracting the Facial Feature Points

Using the open-source software OpenFace, we determined the coordinates and angles of the feature points of a face in a YouTube video [45]. Figure 3.4 shows the image and its feature points when applying OpenFace to a face image. OpenFace converts the face image into 68 feature points as shown in Figure 3.4 and three components of the face angle: pitch (Pose Rx), yaw (Pose Ry), and roll (Pose Rz).

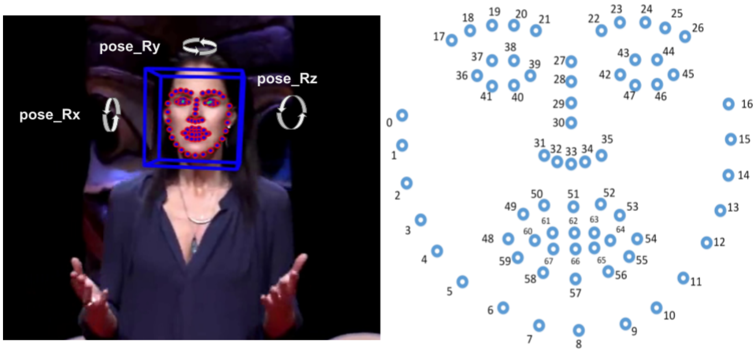

Figure 3.4: The output of OpenFace [46]

3.2.4 Extracting the Body Pose Feature Points

Using the open-source software OpenPose, we obtain the coordinates of the posture feature points of a YouTube video. Figure 3.5 shows the image and its feature points when applying OpenPose to the posture image. When using OpenPose, the output is in json format, so we convert it to csv file [47].

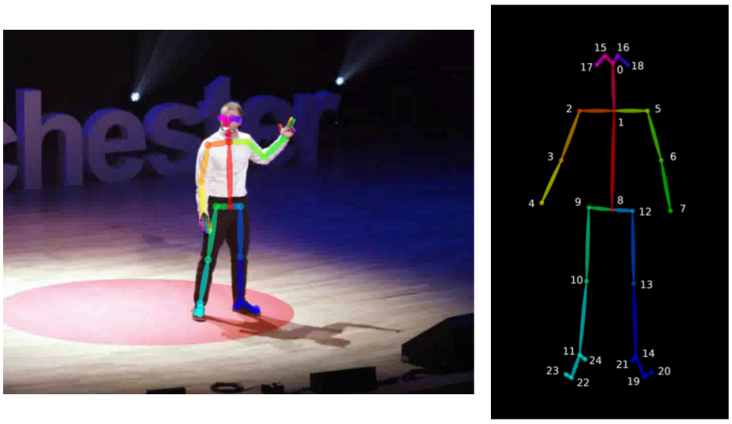

Figure 3.5: The output of OpenPose

3.3 Implementation

3.3.1 Implementation of Classification Models

We create a 7-class classification and a 4-class classification model of the user's emotional categories. We used the features defined and selected to create the models in this chapter. We used three machine learning methods: Logistic Regression, Random Forest and a 3-layer neural network model. 3-layer neural network using Pytorch is described in the next section. For the feature selection, we reduced the number of features by using their correlation coefficients and assuming they correlated with the target values [48]. We remove rows with success = 0 in the OpenFace output and remove rows with missing values in the OpenPose result.

3.3.2 Implementation of Multi-layer Perceptron (MLP) with PyTorch

To build the models, we used PyTorch, a machine learning library. We used iPython notebooks for development and designed our deep neural network model using a GPU on Google Cloud. Looking at Listing 3.1, we have two hidden layers and one output layer. We give the number of features to the linear layer and specify the number of nodes in the hidden layer as 400. Next, we provide that number for the activation function, ReLU(). This Linear and ReLU() are then built up layer by layer. Finally, we select the 7 or 4 emotional categories we want to classify. We performed each model for 100, 300 and 600 epochs, after which the performance did not seem to improve much. Note that the 600 iterations took between 300 and 360 minutes, so it was impossible to iterate while testing the parameters and architecture.

We perform mini-batch training: Train_dataloader contains a set of 32 frames of feature and teacher data. We resize the tensor to feed the linear layer in the mini-batch loop. We use the GPU for training. The idea is to reduce the loss function value by changing the vector value of the weights in the neural network using the backpropagation method. The loss function represents how well the model performs after each optimisation iteration on the training set. We use the Classification Cross-Entropy loss function, and SDG optimised for such classifications. In the optimiser, the learning rate (LR) sets the extent to which the network's weights adjust to the gradient of the loss. Here we set this as 0.001, and the lower this is set, the slower the training will be. To calculate the cross-entropy error, we need to convert the neural network's output to a probability, and PyTorch does this by using the softmax function. We use the softmax function nn.log_softmax to restore the work to a chance [49].

Listing 3.2: A crucial algorithm for the project

```
# This is a class of MLP
class MulticlassClassification(nn.Module):
    def __init__(self, num_feature, num_class):
        super(MulticlassClassification, self).__init__()
        self.fc1 = nn.Linear(num_feature, 400)
        self.fc2 = nn.Linear(400, 200)
        self.fc3 = nn.Linear(200, num_class)

    def forward(self, x):
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return F.log_softmax(x, dim=1)
```

Chapter 4

Testing and Evaluation

In this chapter, we evaluate the performance of the models for estimating the sentiment categories built using machine learning methods on various metrics. In addition, we check the prediction accuracy for each emotion category for the best performing models. Finally, we critically evaluate our results and discuss their shortcomings.

The classification task in this project aims to estimate the user's emotions in communication from externally observable non-verbal information such as facial expressions and body posture. At the beginning of the implementation of this project in a suitable development environment, the distribution of the target variables in the training data showed a significant bias. One approach to imbalanced data here is to see what happens to the predictions when the sentiment classifier is changed from Ekman's level to Group's level to reduce the bias in the distribution of the target variable.

4.1 Evaluation Strategy

We start by observing the balance of the classes. As shown in figure 3.2 and figure 3.3 of 3.2.2 Conducting Sentiment Analysis in Chapter 3, approximately more than 65% of the objective variables in the data indicate the emotion category neutral, which represents an imbalance in the class. Imbalanced data refers to data with a significant bias in the distribution of the objective variable. We should achieve approximately 65% accuracy by simply predicting the emotion category as "neutral" on a test dataset generated from the same distribution as the training dataset. We, therefore, set a baseline prediction accuracy of 65% for this project. For this reason, we did not intend that only Accuracy should be used as an evaluation metric if one of the variable classes covered by the data is in the majority. In the classification problem of this project, where the distribution of the target variable is highly skewed, we will use a measure other than Accuracy, such as Recall or Precision. The criterion for model selection is to ensure that the Accuracy is as high as possible. Further observations using Precision, Recall and F1 scores will take place on the models with the highest Accuracy at Ekman's and Group's emotion classification levels.

We perform the following evaluation studies.

- Model selection using Accuracy
- Training and evaluation of the model with Ekman-level emotion categories
- Training and evaluation of the model at the Group level of emotion classification

The dataset is divided into training, validation and test in the ratio 72:8:20. All evaluations were done on the test data set.

4.2 Model Selection

4.2.1 Emotional Taxonomy: Ekman level

This project uses Accuracy as an evaluation metric for model selection to choose the best algorithm to estimate users' emotions.

Table 4.1: Prediction Accuracy by Model

Model	Accuracy				
	Logistic Regression()	Random Forest()	3 layers Neural Network		
			Epoch100	Epoch300	Epoch600
OpenFace	61.45%	67.78%	67.05%	72.69%	77.04%
OpenPose	59.93%	72.33%	67.28%	70.03%	67.81%
OpenFace + OpenPose	62.27%	79.52%	79.77%	82.04%	82.01%

We can confirm that the emotion estimation model using features from OpenFace+OpenPose is more accurate than those using features from OpenFace and OpenPose alone. In comparing machine learning libraries, we found that Pytorch's 3 Layers Neural Network improved prediction accuracy compared to LogisticRegression and RandomForest. Furthermore, among the 3 Layers Neural Networks, the model with 300 Epochs achieved the highest prediction score of 82.04%.

4.2.2 Emotional Taxonomy: Group level

Similarly to the Ekman level model selection, we performed a Group level model selection using Accuracy as an evaluation metric to select the best algorithm. The results show that the emotion estimation model using the features from OpenFace+OpenPose is more accurate than the model using only OpenFace and OpenPose features. RandomForest has the highest accuracy score of 96.32% compared to the machine learning libraries.

Table 4.2: Prediction Accuracy by Model

Model	Accuracy				
	Logistic Regression()	Random Forest()	3 layers Neural Network		
			Epoch100	Epoch300	Epoch600
OpenFace	79.62%	92.21%	81.29%	85.34%	87.06%
OpenPose	78.86%	94.60%	79.16%	82.91%	85.15%
OpenFace + OpenPose	80.14%	96.32%	91.52%	93.62%	93.40%

4.3 Results

4.3.1 Emotional Taxonomy: Ekman level

For the 3 Layers Neural Network model with an accuracy of 82%, we take a closer look using other measures: precision, recall and F1 for each emotion category classified by Ekman's level, as shown in Table 4.3. The highest precision and recall class is a surprise, with 92% and 93%, respectively. When the user says something surprising, it is more likely to be expressed in non-verbal information. On the other hand, the lowest joy precision and recall scores were 71% and 56%. We can expect that even if a user is enjoying the content of a statement and cannot simply tie it to what expresses itself as non-verbal information. For the other emotional categories (Angry, Disgust, Fear, Sadness, Neutral), the F1 score was around 80%.

Table 4.3: Evaluation metrics (precision, recall, F1) for each emotional category

Emotion Categories	precision	recall	F1	accuracy
anger	81%	80%	80%	
disgust	89%	75%	81%	
fear	87%	79%	82%	
joy	71%	56%	63%	82%
sadness	76%	83%	79%	
surprise	92%	93%	92%	
neutral	84%	90%	87%	

The mixed matrix by emotion shows that neutral accounts for a large proportion of the mixed matrix in this task. The percentage of results where the label joy predicted to be neutral using the prediction model was 9.5% of the total.

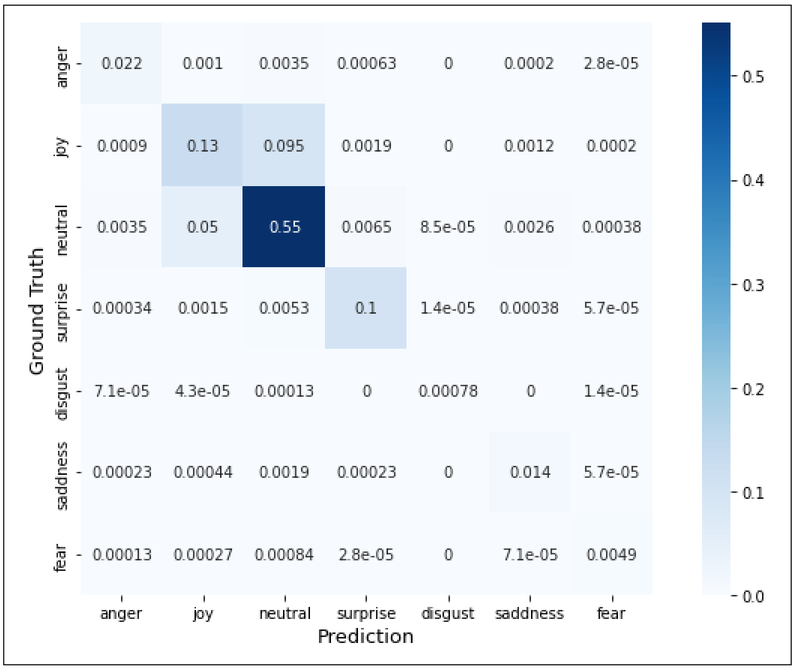

Figure 4.1: Confusion Matrix (Ekman level)

Figure 4.2 shows the number of epochs and the error rate of the OpenFace+OpenPose Pytorch model during training and evaluation. We can see that the error rate of the training model decreases with each epoch. Similarly, the error rate of the evaluation model is decreasing, but we can see some spikes. Although spikes are present, the frequency of spikes is decreasing, which suggests that the learning process is going well.

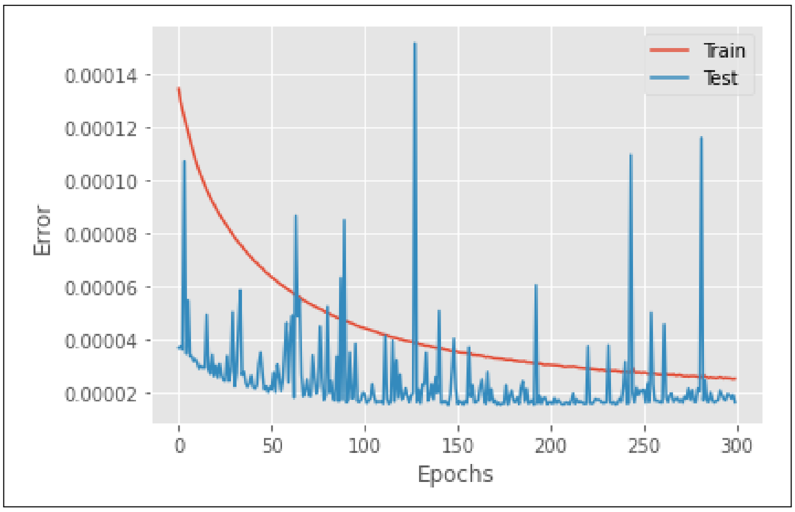

Figure 4.2: Loss function

4.3.2 Emotional Taxonomy: Group level

For the RandomForest model, which had an accuracy of 96%, we will take a closer look using other metrics. Table 4.4 shows the precision, recall and F1 for each emotional category classified by the level of the Group. The highest precision and recall scores occurred in the neutral type, with 96% and 100%, respectively. The proportion of positive, negative and ambiguous precisions was 100%. On the other hand, the negative recall score was the lowest at 79%. We can expect that even if a user says something negative, it may not simply express itself as non-verbal information. For the other emotional categories (positive, ambiguous), recall recorded 82% for both.

Table 4.4: Evaluation metrics (precision, recall, F1) for each emotional category

Emotion Categories	precision	recall	F1	accuracy
positive	100%	82%	90%	96%
negative	100%	79%	88%	
ambiguous	100%	82%	90%	
neutral	96%	100%	98%	

The confusion matrix by emotion shows that neutral accounts for 80% of the total in this task, with 1% expecting something negative to be neutral.

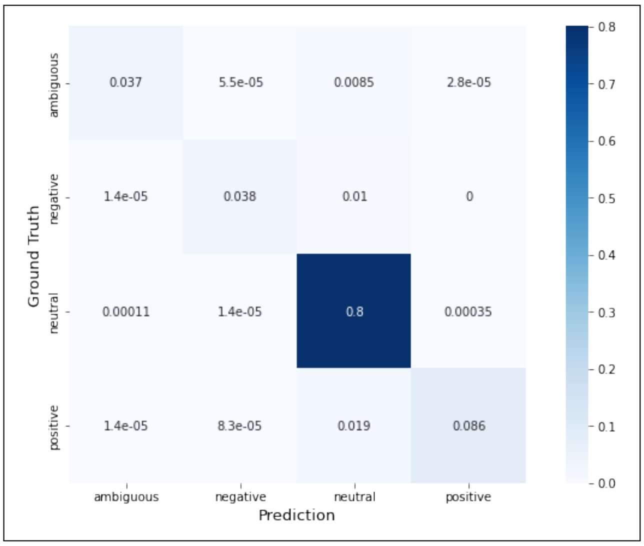

Figure 4.3: Confusion Matrix (Group level)

4.4 Review

This project aims to estimate the user's emotions from externally observable non-verbal information such as facial expressions and body posture in communication. The model's Accuracy created in this project exceeded 65%, which is the baseline of prediction accuracy for the percentage of neutrals in the data set. In addition, we observed that multimodal emotion estimation models that combined face and posture features had overall higher prediction accuracy than models that used only face or posture features. We obtained a higher prediction accuracy by changing the sentiment classification from the Ekman level to the Group level. Group level has a higher percentage of neutral in the dataset, and the change from 7 to 4 classes reduces the difficulty of the classification task. Looking at the machine learning libraries, Pytorch's 3 layers NN model had the best prediction accuracy at the Ekman level, while RandomForest had the best prediction accuracy at the Group level. We expect the system's emotion prediction to be less likely to misunderstand users' emotions by improving Recall. Efforts to strengthen Recall should be considered in the future to prevent miscommunication between users and virtual agents. The figure below shows the output value of log_softmax as the confidence value on the vertical axis and the frame on the horizontal axis to investigate the typical attitude of each emotional category. For example, we have obtained the posture when the confidentiality value of anger is high. In addition, we describe representative examples for other emotional categories in the appendix B.

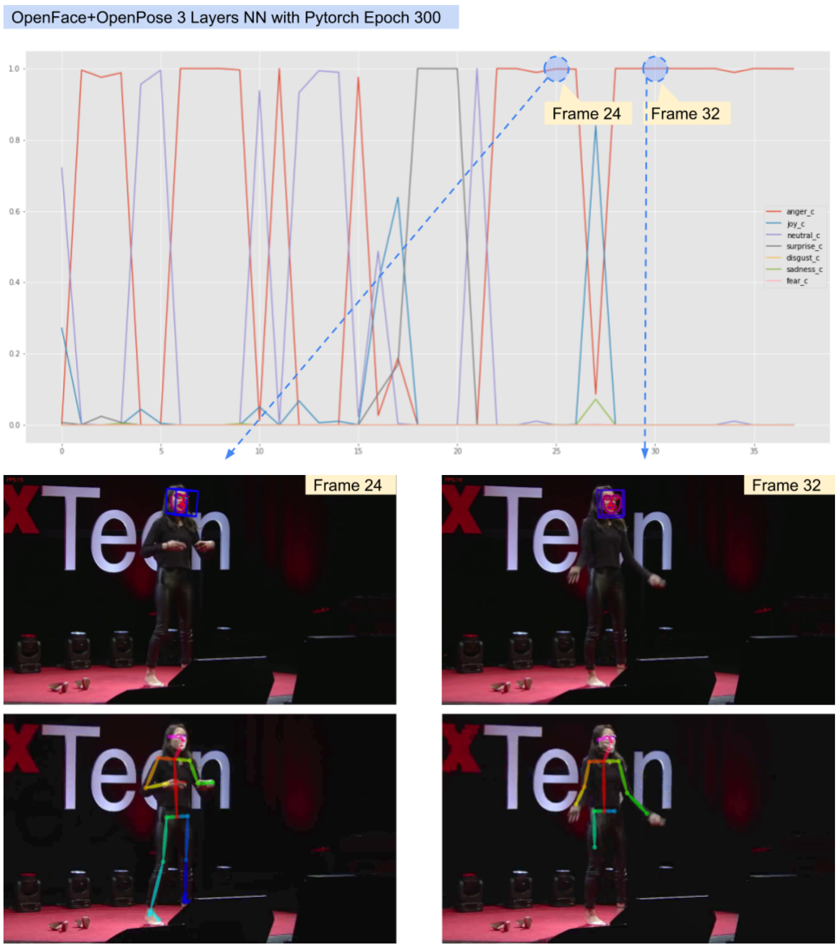

Figure 4.4: Confidence values of the model at each frame and the results of the OpenFace and OpenPose plots

In this project, both training and test data were scarce, and there was a bias towards different emotion categories. However, the proposed model with features on the face, mouth, hands and posture performed better than with only one modality. With more data, we believe that this model could quickly learn appropriate emotion categories from a wide range of non-verbal information, making it a very useful communicative emotion predictor.

Chapter 5

Conclusion

This final chapter discusses the overall review of the study. Section 5.1 discusses the goals achieved and the challenges faced, and section 5.2 examines possible future perspectives for this model.

5.1 Discussion

This study proposed a model for predicting the user's emotions based on externally observable non-verbal information such as the face, hands, mouth, and posture in communication using YouTube videos of people talking to support application users with intelligent agents. We then conducted experiments to evaluate these methods using a variety of machine learning libraries.

The performance of the proposed model was 0.81 for Accuracy at the Ekman level and 0.96 for Accuracy at the Group level. The precision and recall values for surprise were higher than those for the other emotional categories suggests that people are more likely to express surprise as non-verbal information when expressing surprise. The extent to which people show surprise is less biased by individual differences. Despite the limited amount of non-verbal communication in the dataset used in this project, we believe that the model has an excellent potential to predict users' emotions.

In developing the emotion prediction model described above, we collected non-verbal data from videos where the user was talking. In future projects, we could use the same method for other videos to obtain non-verbal information about the user's face, hands, mouth, and posture.

Although the proposed method focuses on frame units, the amount of change of nonverbal information during the user's utterance may also be an essential factor. Therefore, we may improve the proposed approach by obtaining the amount of motion change from the features of non-verbal information obtained from the user and weighting them according to the utterance time. Thus, it is necessary to investigate whether the extension of the model contributes to the improvement of emotion estimation.

5.2 Future Work

Regarding future work, the data size for this analysis was insufficient (14 participants, 3761 sentences). Therefore, we need to reduce the bias in the behaviour of the non-verbal information due to the speakers' nationality, age, and gender and to evaluate the model using videos of other types of discussions. In addition, to ensure an even distribution of the emotional categories, we need additional data for the anger, sadness, fear and disgust classes, which were missing from the current dataset. If this model is helpful for other types of discussions, it will help to suggest the model's versatility. Secondly, to confirm the usefulness of the proposed model, another possible direction is a system that detects the speaker's personality traits and other traits that the speaker is not aware of, such as the way they speak. It suggests a manner of speaking that gives a better impression.

As for the long-term problems, we can anticipate a discrepancy between the actual emotions of the users and the sentiment categories because, in this analysis, sentiment analysis is carried out based on what the users are saying (subtitles) to obtain the sentiment categories. Therefore, it is necessary to get a more accurate objective function by asking the speaker to annotate their feelings or how the listener feels about the statement. Although we obtained non-verbal features from the user, we need to know more about which specific features significantly impact the emotional categories when predicting using the model. We believe that by understanding these characteristics, we can clarify the relationship between emotions and unconscious facial expressions and postures, which will lead to the creation of better virtual agents. Our models and analyses need to be further improved to consider the influence of the face, hands, mouth and posture on the number of non-verbal information features. In addition, although this study focused on the speaker, we believe that research should also focus on the listener, as it is not only the speaker but also the listener's responses and reactions that are important for engagement to occur.

Finally, it is also essential to develop modalities not used in this project to be integrated into the model. At the moment, we do not incorporate all sensor data. For example, the use of eye-tracking data or Kinect sensor data to obtain additional features would enrich the representation of the model and improve the prediction accuracy of the emotion categories. The model could also be shared with other researchers and used as a basis for multimodal interaction research.

Appendix A

Video List

Table A.1: Video list

No	Video Name/ URL
1	Why I Don't Use A Smart Phone — Ann Makosinski — TEDxTeen https://www.youtube.com/watch?v=TjaM0tdxtYA
2	How to speak so that people want to listen — Julian Treasure https://www.youtube.com/watch?v=eIho2S0ZahI
3	How to spot a liar — Pamela Meyer https://www.youtube.com/watch?v=P_6vDLq64gE
4	Robert Waldinger: What makes a good life? Lessons from the longest study on happiness — TED https://www.youtube.com/watch?v=8KkKuTCFvzI
5	How to make stress your friend — Kelly McGonigal https://www.youtube.com/watch?v=RcGyVTAoXEU
6	English Conversation 01 https://www.youtube.com/watch?v=m1-Bx3h4cio
7	How to Be Happy Every Day: It Will Change the World — Jacqueline Way — TEDxStanleyPark https://www.youtube.com/watch?v=78nsxRxbf4w
8	How to Talk Like a Native Speaker — Marc Green — TEDxHeidelberg https://www.youtube.com/watch?v=Ti_gFEe1XNY
9	Why we get mad – and why it's healthy — Ryan Martin https://www.youtube.com/watch?v=0rAngiiXBAc
10	Why you should define your fears instead of your goals —Tim Ferriss https://www.youtube.com/watch?v=5J6jAC6XxAI
11	Ellen DeGeneres' 86th Oscars Opening https://www.youtube.com/watch?v=HUmX6CiMoFk
12	Mathematics is the sense you never knew you had —Eddie Woo https://www.youtube.com/watch?v=PXwStduNw14
13	How to Get Your Brain to Focus — Chris Bailey — TEDxManchester https://www.youtube.com/watch?v=Hu4Yvq-g7_Y
14	How I Overcame My Fear of Public Speaking https://www.youtube.com/watch?v=80UVjkcxGmA

Appendix B

Typical examples

B.1 Surprise

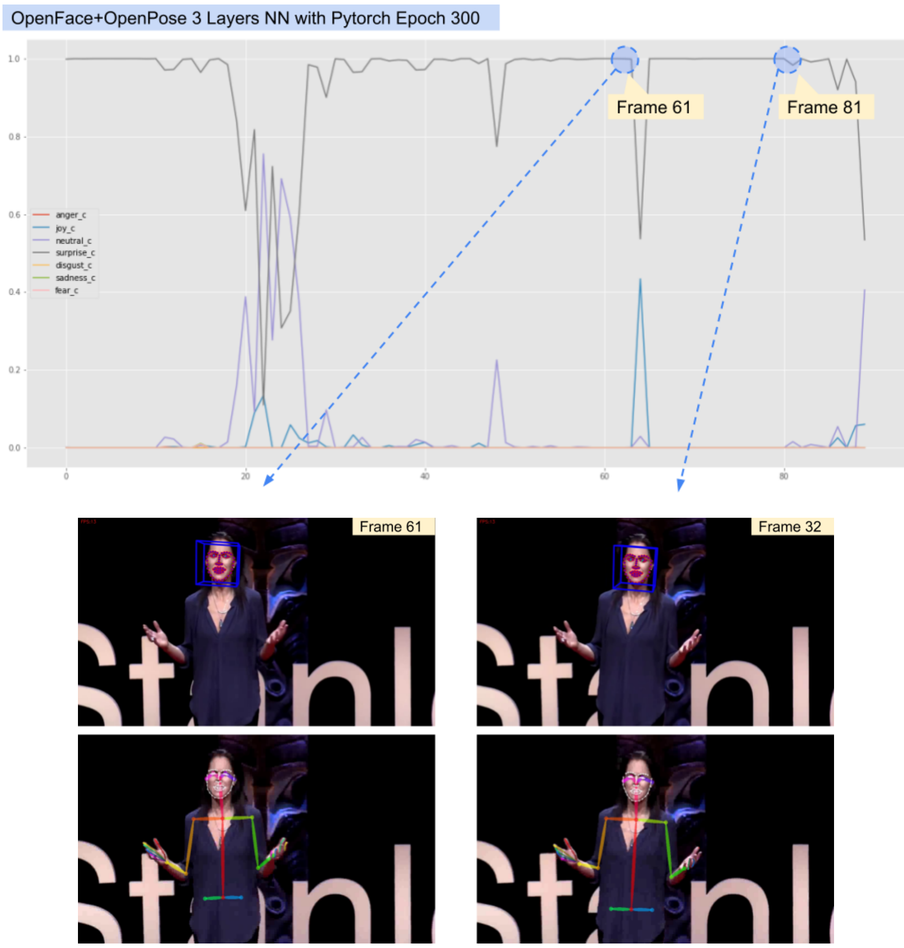

Figure B.1: Confidence values of the model at each frame and the results of the OpenFace and OpenPose plots

Bibliography

[1] M. L. Knapp, J. A. Hall, and T. G. Horgan, *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[2] C. Team, "Five things we learned from releasing a ludens virtual human agent at a shopping mall." Online, 2020. `https://medium.com/couger-blog/five-things-we-learned-from-releasing-a-ludens-virtual-human-agent-at-a-shopping-mall-bf9ba0553751`.

[3] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh, "Where to look: a study of human-robot engagement," in *Proceedings of the 9th international conference on Intelligent user interfaces*, pp. 78–84, 2004.

[4] Y. Yacoob and L. Davis, *Computing spatio-temporal representations of human faces*. PhD thesis, research directed by Dept. of Computer Science.University of Maryland at College Park, 1994.

[5] N. B.-B. Andrea Kleinsmith, Tsuyoshi Fushimi, "An incremental and interactive affective posture recognition system," *International Workshop on Adapting the Interaction Style to Affective Factors*, pp. 378–387, 2005.

[6] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, 2005.

[7] A. Pentland, *Honest signals: how they shape our world*. MIT press, 2010.

[8] M. Johnston, P. R. Cohen, D. McGee, S. Oviatt, J. A. Pittman, and I. Smith, "Unification-based multimodal integration," in *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 281–288, 1997.

[9] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[10] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, vol. 27, p. 46, 1997.

[11] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations.," *Psychological Review*, vol. 99, no. 3, p. 561–565, 1992.

[12] P. Ekman, "Are there basic emotions?," *Psychological Review*, vol. 99, no. 3, p. 550–553, 1992.

[13] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[14] A. J. Calder, A. W. Young, D. I. Perrett, N. L. Etcoff, and D. Rowland, "Categorical perception of morphed facial expressions," *Visual Cognition*, vol. 3, no. 2, pp. 81–118, 1996.

[15] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of the 12th international workshop on semantic evaluation*, pp. 1–17, 2018.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] C.-C. Hsu and L.-W. Ku, "Socialnlp 2018 emotionx challenge overview: Recognizing emotions in dialogues," in *Proceedings of the sixth international workshop on natural language processing for social media*, pp. 27–31, 2018.

[18] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[19] T. Kanade, "Picture processing system by computer complex and recognition of human faces," November 1973.

[20] T. S. Jebara, "3d pose estimation and normalization for face recognition," *Centre for Intelligent Machines, McGill University*, 1995.

[21] R. Jafri and H. Arabnia, "A survey of face recognition techniques," *JIPS*, vol. 5, pp. 41–68, 06 2009.

[22] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[23] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Josa a*, vol. 4, no. 3, pp. 519–524, 1987.

[24] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[25] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[26] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[27] B. Amos, B. Ludwiczuk, M. Satyanarayanan, *et al.*, "Openface: A general-purpose face recognition library with mobile applications," *CMU School of Computer Science*, vol. 6, no. 2, 2016.

[28] K. Mase, "Recognition of facial expression from optical flow," *IEICE TRANSACTIONS on Information and Systems*, vol. 74, no. 10, pp. 3474–3483, 1991.

[29] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Proceedings of IEEE international conference on computer vision*, pp. 374–381, IEEE, 1995.

[30] A. Agrawal and N. Mittal, "Using cnn for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy," *The Visual Computer*, vol. 36, no. 2, pp. 405–412, 2020.

[31] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International journal of computer vision*, vol. 61, no. 1, pp. 55–79, 2005.

[32] L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, pp. 2041–2048, IEEE, 2006.

[33] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, pp. 483–499, Springer, 2016.

[34] Y. Gu, X. Mai, and Y.-J. Luo, "Do bodily expressions compete with facial expressions? time course of integration of emotional signals from the face and the body," *PloS one*, vol. 8, p. e66762, 07 2013.

[35] I. Bartenieff and D. Lewis, *Body movement: Coping with the environment*. Routledge, 2013.

[36] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of nonverbal behavior*, vol. 28, no. 2, pp. 117–139, 2004.

[37] M. Sawada, K. Suda, and M. Ishii, "Expression of emotions in dance: relation between arm movement characteristics and emotion," *Perceptual and motor skills*, vol. 97, p. 697—708, December 2003.

[38] J. M. Murre, R. H. Phaf, and G. Wolters, "Calm: Categorizing and learning module," *Neural Networks*, vol. 5, no. 1, pp. 55–82, 1992.

[39] K. Schindler, L. Van Gool, and B. de Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[41] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10, IEEE, 2016.

[42] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[43] K. Shubham, "Build a youtube downloader with python." Online, 2020. `https://towardsdatascience.com/build-a-youtube-downloader-with-python-8ef2e6915d97`.

[44] J. Park, "Goemotions pytorch." Online, 2021. `https://github.com/monologg/GoEmotions-pytorch`.

[45] M. Kitagawa, "Openface 2.2.0: a facial behavior analysis toolkit." Online, 2019. `https://github.com/TadasBaltrusaitis/OpenFace`.

[46] T. Baltrusaitis, “Output format.” Online, 2019. https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format.

[47] Kim, “Convert openpose json format to csv.” Online, 2019. https://kimbio.info/openpose%E3%81%AEjson%E5%BD%A2%E5%BC%8F%E3%82%92csv%E3%81%AB%E5%A4%89%E6%8F%9B%E3%81%99%E3%82%8B/.

[48] shimopino, “Summary of feature selection.” Online, 2020. https://qiita.com/shimopino/items/5fee7504c7acf044a521#filter-method.

[49] A. Verma, “Pytorch [tabular] —multiclass classification.” Online, 2020. https://towardsdatascience.com/pytorch-tabular-multiclass-classification-9f8211a123ab.