

2023-07

# An AI-based Framework For Parent-child Interaction Analysis

Nikbakhtbideh, Behnam

---

Nikbakhtbideh, B. (2023). An AI-based framework for parent-child interaction analysis (Master's thesis, University of Calgary, Calgary, Canada). Retrieved from <https://prism.ucalgary.ca>.

<https://hdl.handle.net/1880/116757>

*Downloaded from PRISM Repository, University of Calgary*

UNIVERSITY OF CALGARY

An AI-based Framework For Parent-child Interaction Analysis

by

Behnam Nikbakhtbideh

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING

CALGARY, ALBERTA

JULY, 2023

© Behnam Nikbakhtbideh 2023

# Abstract

The quality of parent-child interactions is foundational to children’s social-emotional and cognitive development, as well as their lifelong mental health. The Parent-Child Interaction Teaching Scale (PCITS) is a well-established and effective tool used to measure parent-child interaction quality. It is utilized in both public health settings and basic and applied research studies to identify problem areas within parent-child interactions. However, like other observational measures of parent-child interaction quality, the PCITS can be time-consuming to administer and score, which limits its wider implementation. Therefore, the main objective of this research is to organize a framework for the recognition of behavioural symptoms of the child and parent during interventions.

Based on the literature on interactive parent-child behaviour analysis, we categorized PCITS labels into three modalities: language, audio, and video. Some labels have dyadic actors, while others have a single actor (either the parent or child). In addition, within each modality, there are technical issues, considerations, and limitations in terms of artificial intelligence. Hence, we divided the problem into three modalities, proposed models for each modality, and a solution to combine them.

Firstly, we proposed a model for recognizing action-related labels (video). These labels are interactive and involve two actors: the parent and the child. We conducted a feature extraction algorithm to produce semantic features passed through a feature selection algorithm to extract the most meaningful semantic features from the video. We chose this method due to its lower data requirement compared to other modalities. Also, because of using 2D video files, the proposed feature extraction and selection algorithms are to handle the occlusion

and natural conditions like camera movement,

Secondly, we proposed a model for recognizing language- and audio-related labels. These labels represent a single-actor role for the parent, as children are not yet capable of producing meaningful text in the intervention videos. To develop this model, we conducted research on a similar dataset to utilize transfer learning between two problems. Therefore, the second part of this research is associated with working on this text dataset.

Third, we focused on multi-modal aspects of the work. We conducted experiments to determine how to integrate the prior work into our model. We also provided an ensemble model, which combined the modalities of language and audio based on the semantic and syntactic characteristics of the text. This ensemble model provides a baseline for developing further models with different aspects and modalities.

Finally, we provided a roadmap to support more labels that were not covered in this research due to not reaching enough samples. Our proposed framework includes a labelling system that we developed in the primary stages of the research to gather labelled data. This system also plays a role to be integrated with AI modules to provide auto-recognition of the behavioural labels in parent-child interaction videos to the nurses.

# Acknowledgements

My graduate studies were greatly aided by the unwavering guidance and patience of my supervisor Dr. Moshirpour and my co-supervisor Dr. Duffett-Leger. Their support and feedback proved to be invaluable to this research project.

I am grateful to Dr. Behrouz Far and Dr. Steve Drew for taking the time to review my work and serving on my thesis committee. It is a privilege to have your involvement in my academic journey.

I would like to express my thanks to my parents and family for their assistance and support throughout my studies.

Last but not least, I would like to appreciate my friends and lab mates. They have always helped me to keep going and to stay motivated.

# Table of Contents

<b>Abstract</b>	ii
<b>Acknowledgements</b>	iv
<b>Table of Contents</b>	v
<b>List of Tables</b>	vii
<b>List of Figures</b>	viii
<b>List of Symbols</b>	ix
1 Introduction	1
1.1 Motivation	1
1.2 Research Objectives	3
1.3 Research Contributions	6
1.4 Significance of the research and thesis organization	7
2 A Semantic-based Model for Human Behavior Analysis in Parent-Child Interactions	8
2.1 Abstract	8
2.2 Introduction	9
Data	9
HAR	10
Domain	11
2.3 Literature Review	11
2.3.1 Knowledge-based works	13
handcrafted features	13
hierarchical models	14
rule-based models	14
semantic-based approaches	15
2.3.2 Data-driven works	16
2.4 Methodology	17
2.4.1 Preparation	17
2.4.2 Feature extraction	18
Spatiotemporal features	18
Emotion features	18
Tracking features	18
2.4.3 Semantic features	20
Direction	21
Distance	22
Velocity	22
Emotion	23
2.4.4 Feature selection	24
2.5 Results and Evaluation	24
2.5.1 Dataset	24
2.5.2 Evaluation	25
Data-driven approach	26
Statistics-based approach	26

	Proposed approach . . . . .	27
	2.5.3 Results . . . . .	27
	2.6 Conclusions and Future Work . . . . .	29
	2.7 Acknowledgement . . . . .	29
3	Behaviour Analysis of Parent-Child Interactions from Text . . . . .	32
	3.1 Abstract . . . . .	32
	3.2 Introduction . . . . .	33
	3.3 Related Work . . . . .	39
	3.4 Methodology . . . . .	43
	3.4.1 Dataset . . . . .	43
	3.4.2 Model . . . . .	45
	ML-based model . . . . .	45
	Deep-learning-based model . . . . .	45
	Proposed model . . . . .	46
	3.5 Evaluation, Results, and Discussion . . . . .	48
	3.6 Conclusions . . . . .	54
	3.7 Acknowledgement . . . . .	55
4	A Model for Parent-Child Interactions Analysis from Text and Audio . . . . .	56
	4.1 Abstract . . . . .	56
	4.2 Introduction . . . . .	57
	4.2.1 Semantic characteristics . . . . .	59
	4.2.2 Syntactic characteristics . . . . .	61
	4.3 Related Work . . . . .	62
	4.4 Methodology . . . . .	65
	4.4.1 Dataset . . . . .	65
	4.4.2 Audio classification . . . . .	67
	4.4.3 Text classification . . . . .	68
	Deep-learning-based model . . . . .	68
	Fine-tuning . . . . .	69
	Transfer-learning . . . . .	70
	4.4.4 Combined model . . . . .	70
	4.5 Evaluation and Results . . . . .	73
	4.6 Conclusions . . . . .	75
	4.7 Acknowledgement . . . . .	76
5	Conclusion and future works . . . . .	77
	5.1 Summary and Conclusion . . . . .	77
	5.2 Limitations . . . . .	78
	5.3 Future Work . . . . .	79
	<b>Bibliography</b> . . . . .	80
A	PCITS Labels . . . . .	93
B	Video Tracking Feature Extraction Algorithm . . . . .	95
C	Labelling System . . . . .	96
D	Copyright Permissions . . . . .	98

## List of Tables

2.1	Description of the video dataset labels . . . . .	25
2.2	Selected features for the proposed model in video classification . . . . .	28
3.1	PCIT text dataset examples . . . . .	35
3.2	Comparison of evaluation performance for the text model . . . . .	52
4.1	PCITS text dataset examples . . . . .	58
4.2	Text-audio performance results . . . . .	73
4.3	Overall evaluation performance between models in text-audio modalities . .	74



## List of Figures and Illustrations

2.1	Video Preprocessing Flow . . . . .	17
2.2	Video Classification Performance evaluation . . . . .	31
3.1	PCIT text dataset class distribution . . . . .	44
3.2	Proposed text modality model . . . . .	47
3.3	Performance results for the ML-based model text . . . . .	49
3.4	Performance results for the deep learning-based model RCNN text [1] . . . . .	50
3.5	Performance results for the proposed model-text . . . . .	51
3.6	Explanations for the proposed text model on a true positive sample set . . . . .	54
4.1	Dataset class distribution - audio/text . . . . .	66
4.2	Model architecture for text/audio modalities . . . . .	71
C.1	Labelling system . . . . .	97

# List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
<i>PCITS</i>	Parent-child interaction teaching scale
<i>PDCs</i>	Potent disengagement cues
<i>RTCD</i>	Response to child distress
<i>PCI</i>	Parent-child interactions
<i>HBA</i>	Human behaviour analysis
<i>HAR</i>	Human action recognition
<i>DPICS</i>	Dyadic parent-child interaction coding system
<i>SVM</i>	Support vector machine
<i>HOG</i>	Histogram of oriented gradients
<i>RNN</i>	Recurrent neural network
<i>CNN</i>	Convolutional neural network
<i>FCN</i>	Fully connected network
<i>SGD</i>	Stochastic gradient descent
<i>PCIT</i>	Parent-child interaction therapy
<i>NLP</i>	Natural language processing
<i>ML</i>	Machine learning
<i>POS tagging</i>	Part-of-speech tagging
<i>BOW</i>	Bag-of-words
<i>AWS</i>	Amazon web services

# Chapter 1

## Introduction

### 1.1 Motivation

A principal part of the parent-child interaction teaching scale (PCITS) is detecting behavioural potent disengagement cues (PDCs) and response to child distress (RTCD) in parent-child interactions (PCI) [2]. However, labelling of PDCs and RTCDs is time-consuming and error-prone because it needs to meet multiple conditions before assigning a label to the interaction. These symptoms were collected from videos captured by another system named VID-KIDS. Nurses (or coders) attend a three-day training and then pass a reliability assessment. Providers are required to reach 85% reliability and researchers are required to reach 90% reliability. Then they detect the symptoms and provide the required feedback to the users (parents) based on the PCITS.

The main motivation came from a question from Dr. Mohammad Moshirpour (research supervisor) sparked the idea for the study: Is it possible to have a system that stores the actual detected symptoms from the coders, and use machine learning to automate this process? During studies and meetings, we investigated multiple aspects of the problem and performed a feasibility study to draw a roadmap for the project. We first developed a labelling system that helps nurses and coders to enter the detected symptoms into the system. While the labelling system manages the nursing team with multiple coders to maintain the data, after a while, it enabled us with enough samples to evaluate the outcomes of the research with

actual data.

We analyzed the symptoms (now called “labels”) based on their definition and samples. In terms of machine-learning or deep-learning algorithms, the labels are defined in three different modalities of language, audio, and video. For a number of labels, more than one modality is important, and for each modality, a number of limitations and considerations exist.

For video modality, the nearest research topic is human behaviour analysis (HBA) which is a branch of human action recognition (HAR) as a popular video classification task. HBA has a variety of applications, including monitoring the health of elderly people [3], depression detection [4] and measuring engagement [5]. According to the literature in HBA, we found a number of difficulties specific to the domain that we mentioned in chapter 2. We provided some solutions for these challenges.

For language modality, we found similar work in [6] that prepared a dataset in parent-child interaction. We performed analysis on this dataset and proposed a deep learning model to improve the classification performance compared with the original paper. The main motivation for this part was to use the created model in our domain, in a transfer-learning approach, as described in chapters 3 and 4.

For audio modality, we considered labels that are related to the “parent” actor. This assumption helped us to propose an ensemble model to combine the language and audio modalities because the “child” actor cannot produce meaningful text and the language modality is specific to the “actor” parent. The multi-modal ensemble model also combined multiple aspects related to language and audio processing that we described in chapter 4. The multi-modal model could be extended in future works to create an integrated framework for all

the labels from different modalities, aspects, and considerations.

Appendix A shows a number of these labels that we used in this research. Although these labels are not thorough, we selected those based on having enough samples required for implementing machine learning algorithms. Hence, in terms of modalities, actors, and domain complexities, the proposed model can be extended to support the recognition of all other labels (nearly 50).

## 1.2 Research Objectives

The primary goal of this research is the recognition of the PCITS symptoms by using machine-learning or deep-learning techniques. The dilemma here is if there exists a unique solution to cover all the labels, or if the problem needs to be broken by simpler ones. The straightforward way to break this classification task is based on modalities. Therefore, I will answer the following research question:

RQ1: How to develop data preparation, classification algorithms, and evaluation for each modality of language, audio and video? This research question itself could be broken into three research questions:

RQ2: How to classify samples based on the video modality?

For this research question, we focused on PDCs, where the “child” is the main actor in interaction with the “parent”, and meaning is conveyed through body movements, objects, and facial expressions. However, video-based analysis in parent-child interactions poses several challenges, including the complexity and size of video data, the absence of required datasets for indoor activities [7], the lack of depth data in 2D RGB-based videos, and oc-

clusions and inter- and intra-class variability of actions inherited from HAR. Therefore, a knowledge-based approach is necessary for behaviour analysis in PDCs because of the lack of enough samples to cover complexities.

The main objective of this research question was to provide a solution to address the challenges related to 2D inputs and occlusion in body pose features. Our solution involved developing a feature extraction algorithm and a deep-learning-based model with a knowledge-based approach. The feature extraction algorithm aimed to reduce the impact of camera movement that results in changing body position or changing viewpoint in consequent frames. We achieved this by including stabilization to smooth viewpoint changes and normalization to reduce position variance. This problem especially gets worse in scenarios with 2D video files that unlike most of the literature in this domain, do not provide depth information (RGBD) that is robust against viewpoint changes.

We also used a semantic-based model to add extra knowledge to the model, which helps to overcome the problems caused by occlusions and data scarcity [8, 9]. In the video, the problem of data scarcity is more challenging because of the higher number of dimensions compared with language and audio. By producing part-based features, the semantic features allow the model to rely on the remaining parts and reduce the risk of occlusion [10]. We developed motion-based semantic features and a feature selection process to identify simple or atomic activities that are more correlated with the target behaviour. Finally, a classification algorithm maps the atomic activities to the target label.

In summary, the goal of this research question was to develop a methodology that relies on knowledge-based models to create a semantic-based model for body parts, which is crucial in overcoming the challenge of occlusion and simplifying the analysis process when data is

limited.

RQ3: How to classify samples based on the language modality?

Language analysis is a useful tool for evaluating children’s mental health and behavioural development in parent-child interaction therapy. This evaluation can be done by observing free-play or structured tasks between parents and young children [11,12]. In PCITS, certain labels are specifically related to the structured tasks involved in parent-child interactions.

We came across a similar system called “SpecialTime” [6] that contains about 6000 text samples of dyadic interactions between parents and children. The labels in this dataset reflect emotional, semantic, grammatical, and structural characteristics of parent-child interactions during task performance. Since we could not find many studies that use language modality for evaluating parent-child interactions, we chose to use this dataset. The SpecialTime system can help us evaluate our research in comparison with its claimed performance. Furthermore, the dataset was prepared and cleaned directly by trained nurses, making it suitable for our analysis without additional data preparation and cleaning efforts. Additionally, the labels in this dataset are directly related to our defined labels in language modality. So one objective is to use transfer-learning techniques to apply the trained model on this dataset to our case. We fine-tuned a transformer-based model on the dataset and achieved better performance than what was claimed. Moreover, we evaluated the model to demonstrate its ability to recognize the grammatical, emotional, semantic, and structural aspects of the text. Our results indicate that this model can be effectively utilized in our case.

RQ4: How to classify samples based on the language and audio modalities?

In this research question, we aimed to analyze various aspects of language and audio modalities in our complex domain. The challenge with language modality is that labels are

dependent on both semantic and syntactic features. On the other hand, in audio, labels are mostly based on emotional states, which can be modelled as a type of semantic. Also, most labels in language modality are somewhat dependent on audio. To address this, we proposed four classification modules based on accepted assumptions in audio and language processing. One module is responsible for audio recognition of emotional aspects in audio, while the others focus on recognizing semantic and structural aspects of the text.

We proposed the modules by using some accepted assumptions in audio and language processing. We employed CNNs to recognize emotional states in both text and audio [13,14], RNNs to recognize syntactic or structural characteristics of the text [15], and transformer models to recognize both semantic and structural aspects [16]. We also incorporated the trained model from our previous work as one of the modules.

The other objective is to provide an initial solution to combine multiple modalities. We addressed this objective through the implementation of an ensemble model. However, due to limitations in sample size and classification performance for all labels, we were unable to incorporate all modalities at this stage. The proposed ensemble model outperformed the performance of each individual four modules and can accurately recognize grammatical, emotional, semantic, and structural meanings in our domain.

### 1.3 Research Contributions

Following is a list of my contributions:

1. Designing, development and deployment of the labelling system. The system currently has 8 users, written in React.js + Node.js, and deployed on AWS.



2. Development of AI modules for each chapter in Python based on Pytorch and Sklearn frameworks.
3. Submitted each chapter of my thesis to an accredited conference or journal.

## 1.4 Significance of the research and thesis organization

The result of this study can be used to cover all 50 labels in PCITS by the time that gradually more labelled data is available. In fact, all of the labels have the same characteristics in terms of the modalities, actors, and other aspects related to each modality. All of the variations between these labels could be handled in the same way that is presented in this research. Finally, an ensemble model will be able to provide an integrated solution for all the labels.

The rest of the thesis is organized as follows: Chapter 2 explained the proposed model for the “video” modality, Chapter 3 discussed the solution to be used in parent-child interaction based on the “language” modality from the “parent” actor, Chapter 4 demonstrates the proposed model for the “audio” and “language” modalities on the “parent” actor, and Chapter 5 provides a conclusion and possible future works.

## Chapter 2

# A Semantic-based Model for Human Behavior Analysis in Parent-Child Interactions

### 2.1 Abstract

The parent-child interaction teaching scale (PCITS) is utilized in basic and applied research studies as well as public health settings to pinpoint problematic areas in parent-child relations. In this research, we focused on dyadic parent-child interactions as a part of human behaviour analysis (HBA). We divided related works in interactive video human behaviour analysis into two broad categories: data-driven and knowledge-based methodologies. Among knowledge-based approaches, employing semantics is an essential method to gain control over interactions' constraints. These constraints are imposed by the complexity and absence of data with precise labelling. The goal of this research is to create a PCIT classification task solution that can also be used to measure engagement and disengagement in interactive scenarios and other similar situations.

The key drawbacks in interactive video-based HBA are a lack of data, the complex nature of dyadic behaviour, and a high percentage of missing values in the joint features of the skeleton resulting from viewpoints. In this research, a method for gathering 2D trajectory features for behaviour analysis was put forth. In addition, we created a brand-new joint tracking algorithm that works in the recognition of human action in special situations like excessive camera motions and improper viewpoints. We utilized a feature selection procedure

for semantic characteristics in the model evaluation. In comparison to the literature, the results demonstrate a significant improvement.

## 2.2 Introduction

HBA is a subfield of human action recognition (HAR) in AI that have a number of uses, such as monitoring elderly patients' health [3], identifying depression [4], and evaluating engagement [5]. One potential use in this subject is finding behavioural potent disengagement cues (PDCs) [17] in parent-child interactions (PCIs) in accordance with the PCITS [2]. Traditionally, researchers or healthcare practitioners have manually coded this measurement. Coders must first complete a three-day training course and then successfully pass a reliability test in order to become dependable. Researchers must achieve 90% reliability, while providers must meet 85%. In this assessment, PDCs are derived from the child and parent's interacting behaviours with reference to body movements, objects, facial expressions, and audio aspects. As a result, in terms of HAR, the parent and the child are regarded two actors in an interactive human activity recognition system.

The literature review identified the following difficulties with video-based analysis of parent-child interactions:

**Data** The primary type of data used in behaviour analysis is video. The analysis in HBA is a difficult task due to the videos' complexity, size, and spatial-temporal structure when compared to images and texts. Additionally, compared to outdoor activities, fewer works have been done that focus on indoor behaviours where behaviour analysis is focused on [7]. Sports and other outside activities typically entail more clearly defined qualities than inside

activities. Thus, there is a dearth of the necessary datasets and research in this area. Many studies are compelled to use a knowledge-based methodology due to a lack of data (e.g., [8,9]). Due to a lack of relevant data, we used additional context semantics in our study.

The other issue is with conventional 2D RGB-based videos that lack depth information. The majority of the literature concentrates on RGBD data that is resistant to viewpoint shifts because it has three-dimensional dimensions. This issue can be solved by estimating a 3D pose from 2D points, but this transformation propagates errors. We discovered that this change does not work in high-occlusion situations, where some body parts, such as the legs, are completely undetectable. In the classification task, a number of characteristics, particularly in engagement behaviour analysis, such as *pitch*, *roll*, and *yaw*, are determinants [18]; however, these features demand possessing 3D body joint points. To lower the probability of viewpoint variant points, we concentrated on 2D coordination in this study by using a few normalization and stabilizing stages. Both normalization and stabilization were used to lessen the impact of positional variation and to smooth shifts in viewpoint.

**HAR** The difficulties faced by HBA are similar to those of HAR. HAR involves various obstacles such as occlusions, changes in light and camera movement, and differences in the actions performed by individuals within and across classes. However, some of these challenges are less significant in indoor HBA. For instance, light changes are not a major issue, and the camera typically moves within a limited range, causing minor shaking. To overcome occlusion-related problems, we used semantic features to recognize body parts based on their absence or presence. This approach allowed for the extraction of motion-based features for each body part, resulting in a model that is less dependent on the entire body and is

therefore less prone to occlusion. Furthermore, this technique helps to overcome problems caused by insufficient data in deep learning models. According to a study by [10], the use of semantics is effective in managing intra-class variability and occlusion-related challenges.

**Domain** The majority of HAR research focuses on simple activities with low similarities between different classes (e.g., sports). In contrast, behaviour analysis is more complicated because it involves interactions between multiple actors. This complexity has led to the adoption of knowledge-based models in HAR research, as they can break down the problem into simpler components [19]. This study presented a feature selection process that identifies simple body part activities that are highly correlated with the target behaviour, such as a pendulum-like movement in the left arm. By using these simple atomic activities, the complex and multi-actor behaviour of the target can be recognized.

In summary, this research aims to develop a knowledge-based approach for creating a semantic-based model of body parts. This approach can overcome the challenge of occlusion and reduce complexity in situations where there is limited data.

## 2.3 Literature Review

Video-based human action recognition (HAR) as a part of human behaviour analysis (HBA) has recently attracted much attention. Although there have been numerous advancements made in this area, it is still a very difficult undertaking. Due to the intricacy, scale, and spatiotemporal nature of videos compared to images, audio, and text, some of these difficulties are intrinsic to this field. Researchers discuss occlusions (such as self-occlusion, occlusion of another object, and partial occlusion; [20]), clutter backgrounds, viewpoint or angle of the

camera, varying light, noise, changes in scale and blur in the video, and inter- and intra-class variability of actions as some of the inherent challenges of HAR in videos [10, 21–23].

In addition, there are several additional difficulties with HAR that are unique to behaviour analysis and are restricted by the scarcity of datasets. However, there are still a few associated works and datasets for indoor activities that are helpful for applications in healthcare [7]. The majority of works in the literature are organized and evaluated based on outdoor activities like sports. The makeup of classes has an impact on HAR’s complexity as well. The majority of the most recent research in this area focuses on straightforward, unrelated activities like *sporting activities*, however, these categories do not apply in real-world healthcare applications.

According to [10], the *action* in HAR can be categorized into four categories: *atomic actions*; *people interactions*; *human-object interactions*; and *group activities*. Meanwhile, according to [19], HAR techniques can be divided into two broad groups: *single layered*; and *hierarchical* approaches. While hierarchical approaches first identify simpler actions (also known as sub-events or atomic actions) and then attempt to identify more complex actions (actions that involve objects, contain a sequence of sub-events, or are interactive) based on the simple events, single-layered methods operate on sequences of images. Hence in HBA, due to their ability to deconstruct large classification problems with little to no labelled data into smaller ones and answer them using additional knowledge, hierarchical models are more trustworthy.

According to the categorization presented in [24], HAR methods can be classified into two main categories: *knowledge-based*; and *data-driven-based*. Data-driven approaches involve end-to-end deep learning models that rely on large datasets, which are now feasible due to

the abundance of online video content and powerful processing capabilities. However, the main drawback of this approach in HBA is that these datasets are typically very general and lack specificity. In contrast, knowledge-based solutions rely on domain knowledge in feature engineering, rules, and behaviour reasoning [25]. The primary challenge for knowledge-based techniques is the variability in user behaviour. On the other hand, data-driven models perform better for complex behaviours but are limited in their ability to learn only observable behaviours [25].

### 2.3.1 Knowledge-based works

Knowledge-based solutions involve using prior knowledge to address the problem at hand. This approach has several advantages, including the fact that it is not reliant on large datasets. Furthermore, this technique can facilitate the learning of more complex categories based on the prior knowledge gained from simpler ones. For instance, [10] suggests that having knowledge of previous activities such as “separate egg” and “prepare onion” would make it easier to recognize the new action of “prepare scrambled eggs.” The literature suggests that knowledge-based techniques can be split into the following categories:

**handcrafted features** Handcrafted features involve using prior knowledge to represent, select, and reduce the spatial and temporal features of actions into a classification model such as the Support Vector Machine (SVM). In handcrafted solutions, the classification task is typically straightforward, with most of the effort being devoted to feature engineering [26]. However, the main drawback of this approach is its limited generalizability. For example, the Histogram of Oriented Gradients (HOG) does not perform well when the actors are not positioned in a straight line relative to the camera. As a result, most of the literature

assumes certain conditions must be met before utilizing these features.

**hierarchical models** The purpose of hierarchical models is to break down complex actions into smaller sub-events or atomic actions. For example, [8,9] aim to recognize higher-level activities by reasoning about sub-events. In [27], the authors discuss a multilevel approach to behaviour analysis, where objects (including human body parts) and gestures (e.g., stretching an arm) are detected at lower levels, and simple actions and complex activities are recognized at higher levels. Hierarchical models are especially useful for complex activities, where there is limited labelled data and high variability in both inter- and intra-class. However, these models may not perform well in certain situations, such as occlusions where specific body parts (e.g., legs, hands) may be obscured and sub-events may not be recognized.

**rule-based models** Hierarchical models that use rules as a way to incorporate prior human knowledge into the problem are another approach. These rules can either be explicitly defined by domain experts (e.g., [25,28]) or learned through labelled data (e.g., [9]). However, there is usually a degree of uncertainty associated with these rules. To address this issue, [29] developed a hierarchical framework that first recognizes simple motions using an unsupervised learning technique and then obtains the rules as a decision tree. In [30], fuzzy logic was used for reasoning and a rule base obtained by FCM clustering performed the recognition. Obtaining the required knowledge from unlabeled data is another solution to prepare the rule base. For example, in [24], an ontology was learned from text-based resources negating the need for labelling. Meanwhile, in [18], sub-events are recognized using a feature selection framework from a large pool of statistics-based features obtained from time-series skeleton data.



**semantic-based approaches** Semantic is something with context-specific meaning. Semantic-based approaches in HAR focus on the meaningful relationships (through scenes, objects, and attributes) between actions and body parts in a given context [10]. According to [10], semantics can be useful for intra-class variability by incorporating additional hidden aspects to the model that might not be detected as spatiotemporal features. Semantics are also robust to changes in the shape of the body, clothes, and viewpoint, and can reduce the impact of occlusion in HAR. Furthermore, the semantics of objects, backgrounds, and scenes can be helpful in applying the human understanding of the activity because some behaviours are tied to particular objects (i.e., [31]). Semantic-based attributes can describe a particular characteristic of an activity and be useful for complex activities in HBA. These attributes are characteristics that describe a specific activity, such as “putting one foot in front of the other” in “walking” or how certain activities depend on the gender, type (e.g. “pendulum-like”, “up-down”) of motion in a body part, and the “speed” of motion. Another advantage of semantic-based attributes is that they can be used when there is not enough labelled data for certain classes. This approach allows the system to recognize new events with little or no training examples and transfer knowledge between classes, as described in [32, 33].

As a result, the knowledge can be based on previous human knowledge or gained by machines, and a variety of techniques are used to extract this knowledge. Overall, knowledge-based approaches can overcome some limitations of HAR, and in this research, several semantic-based features are defined.

### 2.3.2 Data-driven works

The advancements in deep learning have recently gained more attention in HAR [34–36], as shown in various image processing tasks where end-to-end deep learning methods outperform conventional machine learning methods that require a lot of effort in feature manipulation and prior knowledge modelling. For example, the 2012 ImageNet Challenge, [37] demonstrated that deep models like AlexNet, consisting of 12 layers such as convolutional layers, nonlinear activation layers, normalization, pooling, fully connected, and classification layers, reduced error rates by nearly half compared to the previous winner. Deep learning models extract different data representations as input is passed through layers, whereas traditional models require hand-engineered task characteristics, which can be time-consuming and require expertise. However, deep learning methods require a large amount of training data and computing resources, making it difficult to apply to interactive behaviour analysis when data is insufficient.

The use of deep learning models has been gaining attention in the field of interactive behaviour analysis. Recurrent neural networks (RNNs), including LSTM [38] and BRNN [39], have been employed to capture the temporal sequence of moving objects in frames [40–43]. Meanwhile, convolutional neural networks (CNNs) have been utilized to capture spatial representation or behavioural patterns and can reduce dimensionality while retaining key characteristics [26, 44–46]. Graph neural networks (GNN) have also been used to capture the interdependence and correlation between body parts’ movements in human actions, as seen in some recent research [47, 48].

While deep learning has achieved good results in recognizing simple actions in HAR,

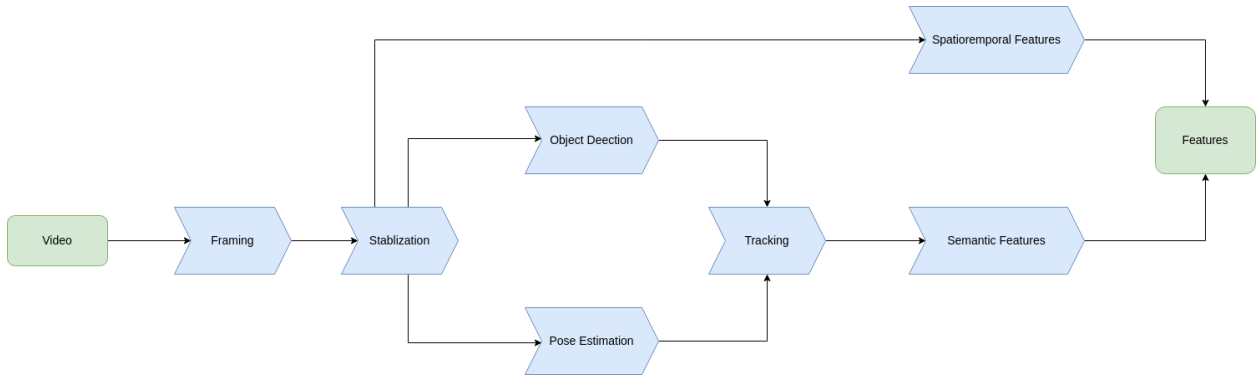


Figure 2.1: Video Preprocessing Flow

applying it to complex human interactions in healthcare is still challenging. [22] discusses the difficulty of using multiple modalities such as RGB, motion, depth, audio, language, and trajectory as features in HAR. The use of depth as a feature requires specialized cameras, as demonstrated in [23], and is not always feasible.

## 2.4 Methodology

### 2.4.1 Preparation

To begin, a data processing pipeline was established, which is depicted in figure 2.1. To prepare the data, videos were sampled at a rate of 20 frames per second utilizing FFmpeg <sup>1</sup>. Next, video stabilization was implemented to eliminate undesired camera movements. A variety of video stabilization techniques are discussed in detail in [49]. We employed a straightforward method that tracks the optical flow between consecutive frames using OpenCV <sup>2</sup> to estimate the motion transformation matrix. The stabilized frames were subsequently fed into two procedures to generate spatiotemporal and trajectory features.

<sup>1</sup><https://ffmpeg.org/>

<sup>2</sup><https://opencv.org/>

### 2.4.2 Feature extraction

Three groups comprise the required features for this study:

**Spatiotemporal features** The video segments contain both spatial and temporal features and vary in length, so to standardize the length, 32 frames from the center of each segment were chosen. Zero padding was applied for shorter segments. The frames were then resized to  $112 \times 112$  pixels and mean normalization was performed by subtracting the mean of R, G, and B from each channel, following the method recommended in [50].

**Emotion features** DeepFace [51] was utilized to detect emotions from the facial expressions of both the child and parent. The emotions are classified into seven categories including *angry*, *disgust*, *fear*, *happy*, *neutral*, *sad*, and *surprise*, and are represented by values between  $[0, 1]$  that indicate the level of confidence. To address the low accuracy of confidences near zero, a filter was implemented to assign a zero value to features below a certain threshold. Based on initial experiments, the threshold value was set to 0.7.

**Tracking features** During this stage, we extracted the 2D body joint locations and objects from every frame. To extract 2D body joint locations, we utilized OpenPose [52], which produced skeleton 2D points. These points are also known as key points, and they consist of 25 joint locations for each human pose, covering the “head”, “torso”, “hands”, and “legs”. In addition, we used YOLOv4 [53] to detect bounding boxes covering people. Both OpenPose and YOLOv4 were employed on each frame. To detect people, we filtered the objects recognized by the pre-trained YOLOv4 model on the COCO dataset.

Because of the significant presence and effect of partial occlusion in our study, where at least one-third of the joints were not identified, the tracking algorithm used was more

intricate than in comparable cases. Simple algorithms have been proposed in various studies such as [18,54] to merge and standardize pose and person information and discard irrelevant data. However, because of the high rate of missing values and false detection, as well as complex factors such as multiple actors in the scene and camera movement, we devised a new tracking algorithm outlined in Appendix B. Empirical experiments showed that this algorithm yielded superior results in comparison to other similar methods.

The algorithm aims to associate skeleton joints or key points with the boundaries of individuals while eliminating partial occlusions. The primary tracking operation was conducted based on clustering the points. A smaller average area cluster of persons was assigned as the child actor, and the more overlapping cluster of key points was assigned as the keypoint cluster of the child. To cluster, the algorithm used *k-means* with  $k = 2$  and Euclidean distance as the similarity metric. The distance was calculated between the vectors of two frames, with the target  $x, y$  points of joints as vectors for key points and the target left, top, right, and bottom points of rectangles as vectors for persons. The similarity distance was averaged between dimensions without considering zero values to disregard the partial occlusion's zero results. Before clustering, the positions of child/parent actors were transformed to the parent's head's origin to remove the camera movement's impact. This was due to the fact that by moving the camera, actors' relative positions remained accurate in succeeding frames when those frames were distinguished by the movement of an actor. Because the parent's head is crucial for extracting semantic features, and the parent's head position had relatively few zero-values in key points, we chose it for the coordinate origin. With a few variations, the final normalization was identical to [55]. The result of the tracking was transformed from the child's key points to the parent's head, then rescaled by the image dimensions,

and quantized to discrete integers in the range  $[0, 100]$ . This was followed by rescaling to the range  $[0, 1]$  for better machine learning analysis that the effect of incorrectly estimated joint coordinates will be lessened by normalization and quantization. The final result was six poselets as defined in [10], representing different portions of the body, which enabled the development of semantic features for each poselet separately. The main advantage of poselets is that they record the essential body parts performing activities independently. In fact, even when other body parts are obscured or improperly detected, activities can still be detected, according to [10]. We were able to separately create the semantic feature for each poselet thanks to this representation.

### 2.4.3 Semantic features

The tracking produces a sequence of 2D points for every actor in the analyzed data, which is normalized over time. This type of output is useful for data-driven methods when sufficient labelled data is accessible. However, due to a lack of adequate data, we created a technique that incorporates domain knowledge into semantic features, which is not reliant on vast datasets. These features can be used for any behaviour recognition problem that involves interaction or multiple actors and has limited labelled data.

To start, we made the assumption that each sample can be represented as a sequence of track key points, denoted by  $T_i$  for  $0 \leq i < N$ , where  $N$  is the segment length (32 frames in our case). Each  $T_i$  is a sequence of x and y key points, denoted by  $(KPC_j, KPP_j)$  for  $0 \leq j < 15$ , representing the child and parent actors. We set the number of joints to 15.

Previous research, such as [18], has introduced features based on 3D key points, attempting to estimate 3D positions in 2D coordinates. However, our investigation found that this

technique did not perform well in cases where many joints were occluded, as it requires knowledge of the positions of all body parts. Thus, these techniques were only effective in limited situations that did not reflect real-world conditions.

Our approach to defining semantic features aimed to capture inherent knowledge. Instead of relying solely on statistical measurements, as in [18], we utilized domain knowledge to establish an initial set of features. We then conducted a feature selection process to identify the most relevant features based on feature categories and poselets, filtering out irrelevant and redundant features. The resulting semantic features are categorized as follows:

**Direction** The movement type can be measured through the direction. Different types of movement, such as “pendulum-like motion” and “up-down motion”, have been described in [10]. The initial direction value for  $KPC_j$  in frame  $F_i$ , denoted as  $IDR_{ij}$ , can have nine possible values of -1, 0, or 1, which represent decreases, no change, or increases in either the x or y direction. To normalize this feature, a new feature called  $MDR_j$  was introduced, which is directly related to the “pendulum-like motion” for each joint. This is calculated using the formula:  $MDR_j = Z_j \times (1 - \frac{\sum_{i=1}^N IDR_{ij}}{N})$ , where  $Z_j$  is the number of non-zero values in  $IDR_{ij}$ , and  $N$  is the number of frames or segment length. The resulting feature value for the poselet is the average of the corresponding non-zero features. Additionally, detecting up-down and left-right motions is accomplished by defining parameters  $dirX_j$  and  $dirY_j$ . These are normalized by the function  $norm(x) = 1 - e^{-\alpha x}$ , where  $\alpha$  is set to  $\frac{3}{Z_j}$ . The resulting  $DIR_j$  values indicate the probability of “left-right” and “up-down” movement, and the total number of semantic direction features is 36. To represent time-series values, statistical functions such as minimum, maximum, average, standard deviations, skewness,

and kurtosis are applied to poselet values and their velocity and acceleration. The total number of statistical features indicating direction is 72.

**Distance** The article [5] discusses various attributes of interactive scenarios such as “turning away” and “approaching”, which can help recognize certain types of behaviour. Relative distance is also an important feature in engagement interactions, as it measures touch and attention. To represent the time-series values in a meaningful way, similar to work in [18], statistical values such as min, max, and variance were computed, and a more meaningful representation of the semantic inside time-series values distribution was performed. The distance was modelled in two dimensions: gradual changes and intensity. Gradual changes are modelled as three categorical values - “Approaching”, “Moving away”, and “Fixed” - for both actors in each target poselet. The intensity of distance is measured using a value transformed by the function  $ID(d) = e^{-\alpha d}$ , where  $\alpha$  is a constant decaying factor equal to 10, and  $d$  is the distance value. The same measurement is applied for internal distances between poselets of the child, including the distance between left-body poselets and right-body poselets, and the distance between body poselets and the head. A total of 104 distance features were calculated. Statistical functions such as min, average, standard deviations, skewness, and kurtosis were applied to the time-series distances of each poselet, resulting in a total of 180 statistical distance features.

**Velocity** The formula to calculate the initial velocity  $IV_{ij}$  of a keypoint  $KPC_{ij}$  in frame  $i$  is given as the distance between the keypoint’s previous and current positions, divided by the time elapsed between the two positions. The resulting values for each joint represent a time series that can be simplified by considering the importance of velocity, both in measuring



engagement and in understanding the order of joint movements between child and parent. To simplify the time series, the velocity values are represented in two dimensions: joints and times. Then, min-max-scale normalization is applied to create two matrices ( $VJ$  and  $VT$ ) that show the normalized values in terms of joints and time. These matrices can be used to understand the activity levels of specific joints relative to others or to compare overall activity levels at different times. Skewness is used to measure the order and intensity of the activity, with the formula

$$V_x = \frac{\text{mean}(x) - \text{median}(x)}{\text{standard\_deviation}(x)}$$

. The resulting skewness values provide insight into the irregularity and frequency of activity in each poselet at different time intervals. The total number of velocity features is 18. For statistical features, three functions (standard deviation, skewness, and kurtosis) are applied to the time-series velocity of each poselet. This results in 36 statistical velocity features.

**Emotion** To model the initial emotion vector for each actor,  $IE_{ij}$  is used, where  $i$  represents the time and  $0 \leq j < 7$  denotes the type of the emotion. The order, intensity, and change of emotions in interactive behaviour analysis are essential. The initial emotion vector is summarized by partitioning the time series for each emotion into three parts, and the average of each partition is calculated as the measure of intensity. The resulting features for each actor are  $E_{jk}$ , where  $0 \leq j < 7$  and  $0 \leq k < 3$ . The total number of emotional features is 42. For statistical features, the time-series emotion vector is subjected to four functions, including maximum, average, skewness, and kurtosis. The resulting statistical emotion features have a total number of 56.

#### 2.4.4 Feature selection

To avoid bias, categorical features were converted to numerical values and then grouped into 32 categories, including feature type (distance, velocity, direction, and emotion); actor type (child, and parent); poselet type (head, torso, legs, and hands). Since feature selection methods can be sensitive to training data, similar to the work in [4], two ensembling techniques were used. The first approach involved combining three different feature selection methods, including the Fisher score, Chi-square, and MRMR. The second approach involved applying the same feature selection method to different partitions of the training data. 20 iterations were run to randomly select 80% of the training data for each experiment, and each method was applied to each feature in each iteration. Features that scored higher than the threshold on at least 80% of the iterations were selected and those that appeared in at least 2/3 of the methods were chosen as final features. The policy applied selected 20% of the features. The total number of semantic features was 200, with 40 output features, while the total number of statistical features was 344, with 70 output features.

## 2.5 Results and Evaluation

### 2.5.1 Dataset

Table 2.1 shows a summary of a subset of the labels used. The description of these labels is included in Appendix A. To address the class imbalance, we employed cropping, shifting, and horizontal flip augmentation to increase the number of samples for each label to 500. Spatial dimension cropping and temporal dimension trimming were applied during segmentation. Stratified k-fold cross-validation with  $k=5$  was used to maintain the class

distribution during training. One effect of shifting (trimming) during segmentation was reducing the likelihood of over-fitting.

Table 2.1: Description of the video dataset labels

<b>Behavior</b>	<i>Number of samples</i>
Back arching	85
Crawling away	70
Maximal lateral gaze aversion	207
Overhand beating movements	313
Pulling away	118
Pushing away	107
Tray pounding	328
Nonverbal Soothing	170
Head Nod	62

### 2.5.2 Evaluation

To assess the effectiveness of our method, we compared it to two other models: a data-driven model that utilizes transfer learning and a machine learning model that uses selected statistical features. The data-driven model uses the state-of-the-art model used in HAR, while the machine learning model focuses on a feature selection process to identify the most suitable statistical features.

**Data-driven approach** To analyze spatiotemporal information in videos, we applied transfer learning using the 3D ResNet-34 model [56], which was pre-trained on the Kinetics dataset [57] consisting of over 300,000 trimmed videos from 400 categories. Similar to how 2D CNNs are applied to image classification, 3D CNNs can recognize spatiotemporal in videos. Since video datasets have smaller scales than image datasets, ResNet architecture was investigated to create a deep learning model for easier training. In our implementation, the 3D CNN extracted features from normalized spatiotemporal features, generating a feature vector of size 512 for each sample. Then, we removed the last layer of the pre-trained 3D ResNet-34 model to add a fully connected network (FCN) classifier with one hidden layer, using cross-entropy as the loss function. We set the learning rate to 0.001, the number of epochs to 100, and ReLU as the activation function. The dropout rate was set to 0.5.

**Statistics-based approach** Based on previous studies in [5,18], we used a similar approach by utilizing statistical features and feature selection to compare results. Initially, we used feature selection to choose 70 out of 344 available features based on the best validation outcome achieved through 5-fold cross-validation. We then constructed a fully connected neural network with two hidden layers and a total of 384,966 neurons. The number of layers and neurons was chosen using a random search. Cross-entropy was employed as the loss function, and the initial learning rate was set to 0.001, which was reduced by a factor of 10 every 20 epochs. The initial values of the learning rate and the number of layers were approximated using a random brute force search and further fine-tuned using Stochastic Gradient Descent (SGD). ReLU was selected as the activation function, and the number of epochs was set to 200. To prevent over-fitting, we used a dropout rate of either 0.5 or 0.2.

**Proposed approach** To start, we utilized feature selection to choose 110 features from a total of 544 semantic and statistics-based features, based on the optimal validation result from 5-fold cross-validation. We wanted to compare the effectiveness of our approach against existing solutions, so we selected a fully connected network over other models such as random forest regression and support vector regression. Each sample was represented as a vector of size 40, and the network was comprised of 2 hidden layers with 512 and 96 neurons, respectively. The number of layers and neurons was determined using a random search. The loss function used is cross-entropy, and the initial learning rate is set at 0.001, which decreases by a factor of 10 every 20 epochs. The learning rate and the number of layers were estimated using a brute-force search and then refined with SGD. The model was trained for 200 epochs using ReLU as the activation function. To prevent over-fitting, the dropout rate was set at either 0.5 or 0.2.

### 2.5.3 Results

To evaluate the proposed model’s effectiveness, a table displaying the selected features is presented in Table 2.2. The table highlights how feature selection uncovers domain knowledge, as each element represents a combination of multiple features that are summarized. For instance, the *distance between child’s head and parent’s head* is a description of  $GD_{head}$  for the child’s actor and is useful in learning certain actions like *Back arching*, making domain knowledge coverage essential.

To evaluate the model’s performance, the dataset was split into train, test, and validation sets with a ratio of 60%, 20%, and 20%, respectively. The training/validation was then performed, with the average train/validation performance shown in Figure 2.2. The

Table 2.2: Selected features for the proposed model in video classification

<i>Feature</i>	<i>Description</i>
$child.GD_{head}, child.ID_{head}$	distance between child's head and parent's head
$child.DIST_{head-legs}$	distance between child's head and child's leg
$child.GD_{hands}, child.ID_{hands}$	distance between child's hand and parent's hand
$child.V_{hands}$	child's hand velocity
$child.V_{legs}$	child's legs velocity
$child.E_{angry}$	child is angry
$child.E_{happy}$	child is happy
$child.MDR_{hands}$	frequent child's hand movement
$parent.DIR_{heads}$	parent's head moving up or down

data-driven model displayed over-fitting as a result of complexities within the domain not represented in spatiotemporal features, with a validation accuracy of about 47%. The validation accuracy of the statistics-based model was 55%, while the proposed model achieved a validation accuracy of approximately 68%, indicating better performance. The test accuracy of the data-driven approach was 37%, and the test accuracy of the statistics-based model was around 52%, while the proposed model demonstrated superior performance, with a test accuracy of around 63%.

## 2.6 Conclusions and Future Work

The task of analyzing parent-child interactions using the PCITS assessment tool falls under HBA and can be approached through two methods: data-driven and knowledge-based. However, analyzing such interactions in 2D video files with multiple actors, occlusion, and camera shakes poses various challenges. To overcome these issues, we developed a tracking algorithm to detect skeleton joints of parent and child actors, and semantic-based features were proposed to handle occlusion and limited labelled data. These features were refined through feature selection to break down complex behaviour into simple activities occurring in body parts. This decomposition allowed the model to be trained in reverse by recognizing actions from semantics in body parts. We evaluated the performance of the proposed model through transfer learning using a data-driven approach and a statistics-based model as per prior research. In the future, the framework could include other modalities like language and voice to enhance performance.

## 2.7 Acknowledgement

We would like to acknowledge the generous support of many organizations that funded the EQUIP project. We thank AMS Health Care for the Fellowship in Artificial Intelligence and Compassion, UCalgary for the VPR Catalyst Grant, Alberta Children’s Hospital Research Institute for the Catalyst Seedling Award, and the Social Sciences and Humanities Research Council (SSHRC) for the Insight Development Grant. These sources supported the preparation, labelling, and analysis of video data codes. We also express our gratitude to CIHR for the source PCITS videos (University of Calgary REB# 16-1811). We extend our

heartfelt thanks to the CIHR study participants that make this work possible. We also thank Lyndsay MacKay, Jason Novick, Jennifer Black, Alexa Toews, Chris Street, Tian Westland, Linnea Davison, Harleen Sanghera, and Carl Dizon.



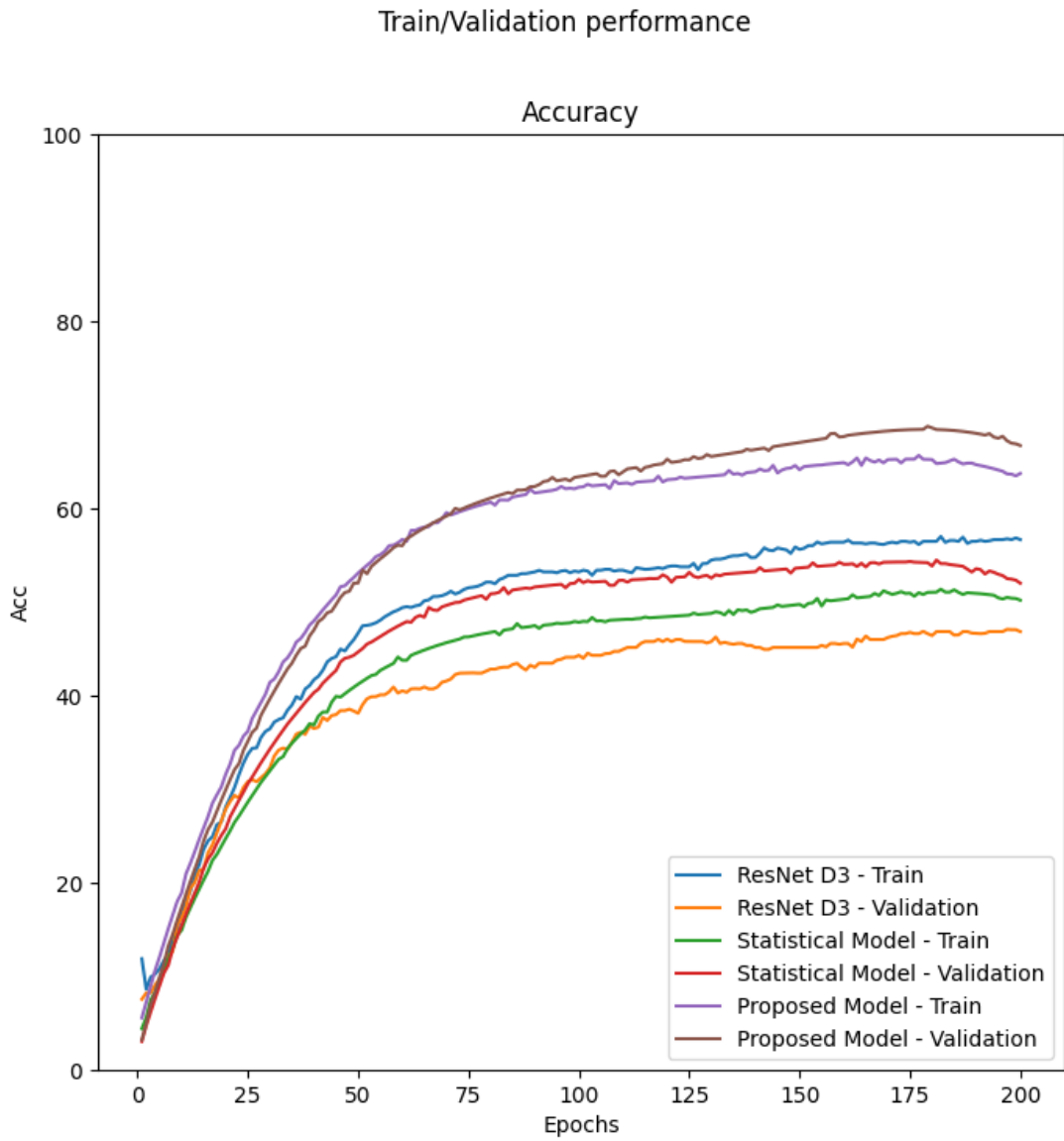


Figure 2.2: Video Classification Performance evaluation

## Chapter 3

# Behaviour Analysis of Parent-Child Interactions from Text

### 3.1 Abstract

Parent-child interaction therapy (PCIT) plays a significant role in determining the trajectory of children’s mental and behavioural health. The quality of interaction between a parent and child can be evaluated by observing unstructured play or guided tasks. Based on these evaluations, healthcare practitioners can implement therapy and offer feedback aimed at improving the quality of these interactions.

However, manual evaluation is a labour-intensive and time-consuming process that often limits its accessibility. This research seeks to harness the power of Artificial Intelligence (AI) to automate some aspects of these therapeutic evaluations, enhancing the scalability and reach of this approach. The goal is to design methods that automatically analyze the quality of interaction based on linguistic elements found in dialogues between parents and children.

A key challenge in PCIT is the classification of parent-child behaviour to assist parents in managing early behavioural issues. In this study, we discuss these facets and challenges with respect to AI. Subsequently, we propose a solution for classifying major behavioural classes in the Dyadic Parent-Child Interaction Coding System (DPICS). To the best of our knowledge, our work is the first to use a Transformer-based architecture to analyze emotions and psychology embedded in parent-child interactions. The model we propose is capable of

understanding and identifying grammatical, syntactic, and emotional features of the language used in these interactions.

We grouped Natural Language Processing (NLP) techniques for grading parent-child interaction quality into three categories: those based on deep learning, machine learning, and transfer learning. Our model follows a transfer learning approach, fine-tuned on a RoBERTa model. The evaluation of this approach showed promising performance enhancements. Our model relies on textual data and has been proven to deliver comparable results without using audio. In terms of performance, the evaluation revealed that our solution outperformed other methods on the same scale. The evaluation also highlighted the model’s capability to identify behavioural aspects of parent-child interaction without requiring additional feature engineering or the incorporation of additional data modalities.

## 3.2 Introduction

Parent-child interaction therapy (PCIT) is a powerful determinant of a child’s mental and behavioural health [11]. The quality of parent-child interaction can be evaluated by observing either free-play or organized tasks between parents and young children. Healthcare professionals, including nurses, social workers, and physicians, can initiate therapy based on these assessments and offer feedback aimed at enhancing parent-child interactions [12]. The therapy teaches parents to adopt effective communication strategies when interacting with their children. Within PCIT, a key skill entails mastering suitable ways of engaging in dialogue with a child. Parents are trained on various communication behaviours to adopt frequently during their interactions with their children (such as *labelled praise*, for instance,

“I appreciate your help”), as well as those to avoid (like *negative talk*, such as “Don’t do that”).

However, therapeutic evaluations are often resource-heavy and time-intensive, which can restrict their scalability and widespread use. Therefore, this study primarily aims to employ AI to automate part of these evaluations by identifying the quality of interaction based on the linguistic elements in a parent-child dialogue. This objective leads to several research questions:

1. How can we educate AI models to effectively gauge the quality of parent-child interactions using the linguistic characteristics present in their conversation, while aligning specifically with the PCIT framework?
2. Can we boost the AI model’s performance beyond the 80% agreement rates amongst therapists as noted in [58], particularly aiming to lower false positives and enhance total accuracy?
3. What is the approach to infusing semantic and syntactic traits from PCIT into the AI model?
4. What is the impact of a child’s dialogues on the final classification outcome?  
Would a model focusing predominantly on parents achieve similar outcomes?

A significant challenge in applying AI to this area is the insufficiency of data, which can be attributed to various factors including ethical constraints. For this study, we selected the dataset from the “SpecialTime” system [6]. This system encompasses 6,022 instances of parent-dialogue acts that have been annotated by therapists. The data were labelled using

the DPICS, a dialogue act classification method used in PCIT [59]. Table 3.1 provides details and a few class examples from this dataset. The label descriptions portray the emotional, semantic, grammatical, and structural characteristics of the context. For instance, *negative talk* could suggest a negative sentiment in the text or highlight a sarcastic or impolite remark. Moreover, *labelled praise* demonstrates positive behaviour towards the *child* concerning a particular subject or task. Other labels like *question* and *reflection* are largely dependent on structural and grammatical features. Therefore, multiple aspects of a typical NLP text classification task contribute to this classification problem.

Table 3.1: PCIT text dataset examples

<b>Class</b>	<b>Description</b>	<b>Example</b>
Neutral Talk (NTA)	Statements that do not explicitly describe or evaluate the child’s present or immediately past behaviour.	(1) Parent: zebras do have stripes. (2) Parent: I have no idea what to do next.
Negative Talk (NT)	Verbal comments expressing disapproval about a child’s characteristics, activities, decisions, or outputs. It also encapsulates sarcastic, rude, or disrespectful speech.	(1) Parent: Don’t put that piece there, please. (1) Child: May I eat ice cream later? (1) Parent: You can count on that happening.
Continued on next page		

**Table 3.1 – continued from previous page**

<b>Class</b>	<b>Description</b>	<b>Example</b>
Reflection (RF)	A statement or phrase that mirrors the child’s verbal expression, maintaining the same meaning.	(1) Child: I did it. (1) Parent: You did it. (2) Child: I want to go home. (2) Parent: You want to go home.
Command (CMD)	Statements where the parent guides the child’s actions. These directives can be either overt or subtle.	(1) Parent: Move closer. (2) Parent: Give your jacket to me.
Behaviour Description (BD)	Impartial, declarative phrases or sentences where the subject is the other person and the verb describes that person’s current or immediately previous observable verbal or non-verbal behaviour.	(1) Parent: You are seated in the chair. (2) Parent: You are driving the car.
Question (QU)	Requests for an answer, but they do not imply that the other person should undertake a specific behaviour.	(1) Parent: What happens to that? (2) Parent: What are you doing? (3) Parent: Would you like some raspberries?
Continued on next page		

**Table 3.1 – continued from previous page**

<b>Class</b>	<b>Description</b>	<b>Example</b>
Unlabelled Praise (UP)	Offers a positive appraisal of the child, a feature of the child, or a non-specific activity, conduct, or output of the child.	(1) Parent: I love you. (2) Parent: Many thanks. (3) Child: I built a tower. (3) Parent: Well done
Labelled Praise (LP)	Gives a positive assessment of a particular trait, output, or behaviour exhibited by the child.	(1) Parent: I appreciate you joining me in my Lego. (2) Parent: I appreciate your cleaning. (3) Parent: You created a lovely drawing.

To categorize this data, we delved into the most recent advancements in NLP. Our exploration resulted in a validation accuracy of 90%, substantially higher than the 79% efficacy noted in [6], and the 80% therapist agreement rates specified in [58]. In implementing our approach, we found that excluding vocal features can actually enhance performance. Our results also indicated that *child* dialogues had minimal impact on the classification outcome, barring one label. This suggests that in this domain, the classification task heavily relies on the *parent* dialogues without the need to factor in dyadic complexities. Even though the dialogues inherently have a dyadic nature, there is a scant correlation between classes and *child* dialogues due to a significant number of missing child statements. Approximately 80% of the samples for classes, excluding *reflection*, lack the child’s statements.

Given the intricate and vague nature of the class definitions, employing conventional machine learning (ML) techniques with an emphasis on feature engineering proves to be challenging. Furthermore, there was not sufficient data to carry out deep learning training from the ground up. Consequently, our research findings illustrated that using either deep learning or ML approaches, the use of transfer learning, stemming from a fine-tuned Transformer-based model, surpassed the results achieved by training on the complete dataset.

In our study, we segregated the applicable techniques into three categories: deep learning-oriented, ML-oriented, and transfer learning-oriented. In the research cited as [6], investigators developed a model that appeared to be more accurate than those based on deep learning. However, employing one-hot vectorization in such models does not ensure their generalizability. There is a susceptibility to both over-fitting and under-fitting when one-hot encoding methods are utilized, such as TFIDF, bag-of-words (BoW), or advanced embedding models like word2vec and Glove, along with lexicon-based solutions using part-of-speech (POS) tagging or linguistic inquiry and word count (LIWC). Moreover, these encoding methods fail to account for the word's position in the text and do not acknowledge any semantic interrelations between categories, making it arduous to generalize the model [60]. The only merit these models hold over transfer learning is their interpretability, which we accomplished in our research through certain visualization models.

However, in this study, aiming for both performance and generalizability, we adopted a transfer learning-based model that superseded rival solutions in terms of results. As far as we are aware, this is the first attempt to tackle this issue using a transfer learning approach, especially in situations where adequate data samples are sparse. Moreover, it is the first time that the powerful capabilities of a cutting-edge Transformer-based algorithm have been



employed to address this intricate problem. The evaluation results of our proposed model demonstrated its ability to discern the emotional and structural aspects of the text essential for classification within the DPICS framework.

The remainder of this paper is structured as follows: Section 3.3 provides a review of the relevant literature that was examined during this research. In Section 3.4, we outline the configuration of our model and describe some additional details and improvements. We then present the results of our experiments in Section 3.5, where we compare our model’s performance against a baseline and discuss the findings. Finally, we draw conclusions in Section 3.6.

### 3.3 Related Work

There is a limited amount of research focusing on the NLP analysis of parent-child interaction. Despite the profound impact of the quality of parent-child interaction on child development, observational studies are costly and time-intensive compared to other non-observational approaches [61].

The studies that do exist either focus on audio or text analysis, or a combination of both, with some leveraging vocal features. For instance, [62] used the volume and tone of the parent and child’s voices to create a model for PCIT. In a similar study targeting depression detection, vocal features such as jitter, energy, and loudness were used [63]. These features were trained and merged with the textual feature analysis.

Most of the existing literature in this area is text-based. For instance, [61] put forth a classifier algorithm to differentiate between mothers who had previously received treatment

for depression and those who had not.

Regarding dyadic interactions, several studies focus solely on the parent’s language, omitting the child’s utterances. In these instances, dyadic features that denote parent-child interaction are disregarded (e.g., [61, 64, 65]). Only a few studies, such as [6], pay attention to the interactive aspects of parent-child dialogues. For instance, determining the class *reflection* as a desirable parental response to a child necessitates an understanding of the whole conversation, not just individual sentences.

From a methodological perspective, several studies propose solutions grounded in empirical methodologies. For instance, systems like TalkBetter [66] and TalkLIME [67] monitor conversational flow and issue warnings upon detecting detrimental linguistic patterns from a parent. When it comes to employing AI methodologies, much of the existing literature falls into two broad categories: models based on ML and those using deep learning. ML-based models utilize traditional ML algorithms with either domain-specific or knowledge-based features, which include lexicon-based and handcrafted features. LIWC [68] is one such feature that offers semantically and syntactically categorized words like “first-person pronouns,” “positive” or “negative emotion” words, “cognitive process” words, and “temporal” words for further analysis and training. LIWC is a program designed to analyze natural language. It employs a dictionary-based approach, grouping words into pre-defined syntax and semantics-related categories and then quantifying their usage. Features of LIWC are noted to be beneficial in identifying behavioural conditions, such as depression detection [61]. The model is built on a Support Vector Machine (SVM) with interpretable LIWC features chosen based on empirical research. As another example, LIWC is used in [69] to support the claim that “positive-feeling speakers” would use more “positive affect phrases”, more “vocabulary”,

and fewer “first-person pronouns”. [70] employed logistic regression for classifying cognitive distortions and showed superior results using ML as compared to deep learning.

Among the works that rely on handcrafted or knowledge-based features, [71,72] used the variety of a parent’s vocabulary related to children’s language abilities to identify different aspects of grammatical and emotional relationships between words. [64] introduced a system named “Captive” to coach parents on providing their children with appropriate linguistic input to reduce the risk of delayed language development. To promote children’s language development, these systems alert parents about factors such as the amount of speech, suitable responses, and lexical diversity. These features are grounded in empirical studies suggesting, for instance, that the diversity in a parent’s word usage is linked to their children’s language skills.

To support healthcare providers in evaluating the quality of parent-child interactions, [6] developed a system named “SpecialTime” that offers feedback to parents about their child’s behaviour. Parents are taught a series of conversation acts in PCIT that they should frequently use when talking to their children (e.g., “That’s nice work” as *labelled praise*) and a separate set that they should avoid (e.g., “You’re bugging me” as *negative talk*). *Labelled praise*, *behaviour description*, and *reflection* are the three ideal dialogue acts that are extremely pertinent when assessing parents’ therapy progress. *Question*, *command*, and *negative talk* are considered unfavorable types of speech. *Unlabelled praise* and *neutral talk* are regarded as inconsequential.

In “SpecialTime”, a threshold check on voice tone is employed to identify questions by calculating the first derivative of the pitch contour of the final 0.5 seconds of the spoken segment. However, we found that solely text-based analysis is sufficient for this task. To

optimize the model, one-hot vectorization with TFIDF text representation, coupled with uni- and bigrams along with POS tags, is utilized. Finally, a linear SVM classifier executes the recognition.

Deep learning-based models leverage the semantic and syntactic attributes of the text within the model. It is commonplace in deep learning-based models to use convolutional neural networks (CNNs) to capture emotion and recurrent neural networks (RNNs) to preserve the structural and temporal aspects of the text [73].

In NLP, deep learning-based models typically employ embeddings that are sturdier than lexicon-based and handcrafted features. In these embeddings, words with similar semantic meanings, like “good” and “better,” are mapped to closely located vectors in the embedding space. However, some researchers argue that these embeddings neglect sentiment. For instance, “good” and “bad” are mapped to nearby vectors [60]. Another challenge with deep learning-based models is their need for vast amounts of data. Consequently, prevalent deep learning-based models in the literature focus on straightforward and standard labels like basic sentiment or emotions on publicly available data (i.e., [13]).

Pre-trained models like BERT [74] are trained on large volumes of unlabelled data, enabling them to learn universal representations of language. As such, using this representation may yield better results in scenarios with limited available data. By fine-tuning the pre-trained BERT model with one additional output layer, modern models for various applications can be created. For instance, [65] describes a system providing guidance to parents of children with unique hearing conditions. These suggestions are generated by a semantic similarity measure that uses BERT for text vector representation. While they did not conduct any training in their work, they utilized the contextual meaning of the text from a

general, pre-trained BERT model.

Moreover, fine-tuning facilitates transfer learning, where knowledge gained from one task can be transferred to a related task. This allows a general-purpose pre-trained model to adapt to a new task-specific model, reducing the data points required to train a model from scratch. In a similar work that concentrated on the structural context of the text, [75] used a combination of contextual representation by BERT, syntactic and semantic categorization by POS tags, and LIWC features. They proposed their model as word classification by fine-tuning BERT, followed by feeding part-of-speech features and LIWC into a logistic regression model for the final classification. Nevertheless, in some instances, fine-tuning BERT can capture the meaning, and the semantic and syntactic structure of the text, negating the need to use it as a feature extractor and combine it with LIWC and POS tags. Fine-tuning pre-trained models can also help mitigate over-fitting, a common issue when training deep-learning models on small datasets.

## 3.4 Methodology

### 3.4.1 Dataset

Fig. 3.1 displays the number of samples in our dataset. To mitigate the effects of class imbalance, we adopted two strategies. For traditional ML-based models, we implemented back-translation [76] as an augmentation technique, which translates sentences into multiple languages before retranslating them back to English. Given the TFIDF vectorization, this augmentation method did not lead to over-fitting. For deep learning-based models, we used a weighting approach, akin to the method used in [77], which assigns higher probabilities

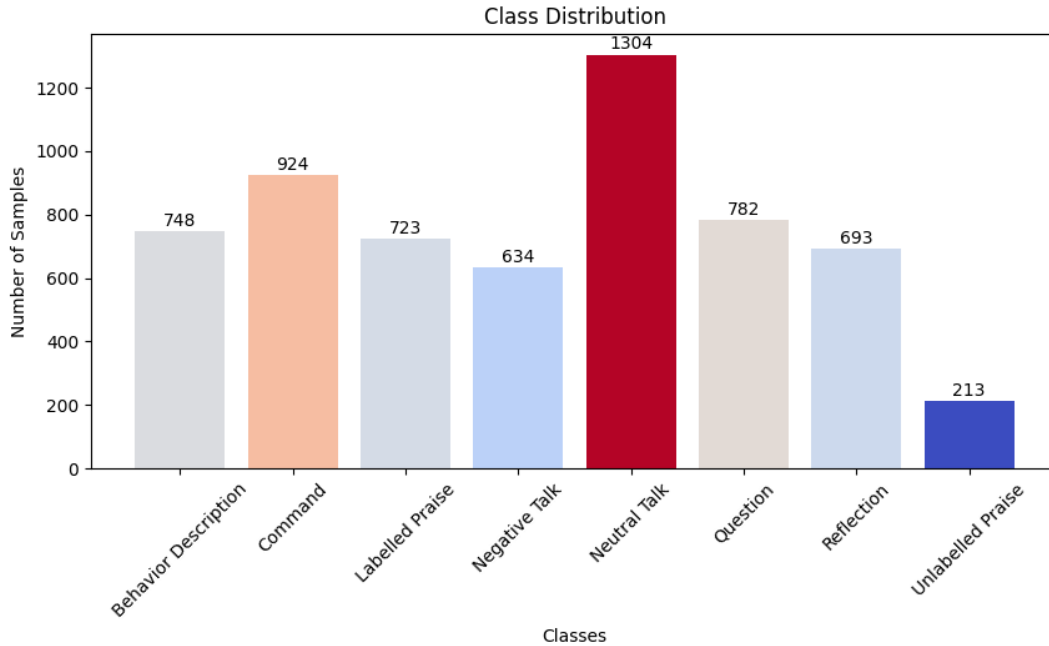


Figure 3.1: PCIT text dataset class distribution

to less common classes in the loss function. However, this weighting can also have negative impacts, such as reducing precision. Therefore, in our final submission, we balanced this during curriculum learning [78], by only employing the weighting strategy in the first half of the training process.

In terms of data preprocessing for ML-based models, we followed common steps used in similar research (e.g., [79]), which include: 1. Removing punctuation: While punctuation aids in structuring text into sentences, its usage in classification can affect the results. 2. Eliminating stop-words: Stop words, like articles and prepositions, do not add any extra meaning. By removing these frequently used words, we can focus on the key terms. 3. Implementing stemming and lemmatization: Stemming reduces words to their root form, for instance, transforming the word “troubles” to its root word “trouble”. In this process, we

used the Porter-Stemmer algorithm. Lemmatization is similar to stemming, but it takes into account the word’s meaning. For deep learning and transfer learning-based models, we only performed the normalization of special characters and punctuation.

Regarding the division of train/test/validation sets, for the ML-based and deep learning-based models, we divided the dataset using a 60% split for the training set, 20% for the test set, and 20% for the validation set. For the proposed model, we split the dataset using an 80% allocation for the training set, and 20% for the validation set.

### 3.4.2 Model

Multiple models were created to find the optimal solution for this problem.

**ML-based model** In accordance with similar research, such as [6, 80], we constructed a traditional ML-based model using an SVM trained on the TFIDF representation of the input text. We also carried out additional experiments to examine the influence of factors such as multi-actor versus single-actor input, k-fold cross-validation, and the use of 1- or 2 grams for tokenization. Ultimately, we found that using 10-fold cross-validation on parent-only texts with 1-gram tokenization yielded the best results. For the tuning of hyperparameters, we performed a 10-fold grid search cross-validation. The optimal parameters were found to be an RBF kernel, with a C value of 1, and a gamma setting of 0.8. The selection of these hyperparameters was based on the accuracy achieved on the validation set.

**Deep-learning-based model** We constructed two deep learning-based models in line with the latest trends for text classification tasks. The decision to use these models was based on the incorporation of CNNs and RNNs, a combination often found in similar research such

as sentiment analysis and depression detection (e.g., [79, 81]).

1. Recurrent Convolutional Neural Networks (RCNNs) [1]: RCNNs, through their recurrent structure, are adept at capturing contextual information and employ a CNN for text representation. This configuration leverages the strengths of both RNN and CNN, creating an efficient and potent model. RCNNs are frequently used in similar research fields, such as sentiment analysis (e.g., [82]).
2. A CNN model following the structure presented in [83]. This model takes word vectors as input and processes them through multiple convolutional layers. These layers employ various filters to identify local dependencies in the data, subsequently generating feature maps. A max-pooling layer is then applied to reduce dimensionality and extract the most significant features. Lastly, these features go through a fully connected layer for classification. We utilized Glove-300 [84], pre-trained on Wikipedia, as the embedding layer with a dimension of 300.

For all deep-learning-based models, we adhered to the architecture outlined in the original references. We set the number of epochs to 15 and the batch size to 32. Cross-entropy was selected as the loss function and AdamW [85], a variant of Adam featuring weight decay, was chosen as the optimizer with a learning rate of 0.0003. To mitigate the potential for over-fitting, we implemented dropout at a rate of 0.2.

**Proposed model** The suggested model we developed employs BERT as a basis for transfer learning. BERT, first introduced by [74], is a bidirectional transformer pre-trained on a vast text corpus, including Wikipedia, using a combined objective of masked language modelling and next-sentence prediction. RoBERTa [86], short for Robustly Optimized BERT Pretraining Approach, is an enhanced language model that improves BERT by adjusting hyperpa-



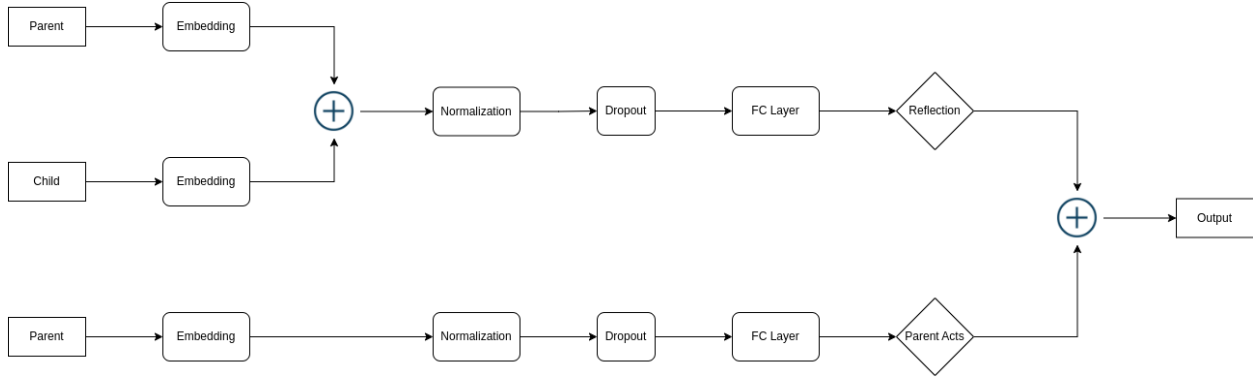


Figure 3.2: Proposed text modality model

rameters and training on a larger dataset. It outperforms BERT in numerous benchmark tasks and serves as the basis for our proposed solution. The successful handling of unstructured data by RoBERTa, combined with its contextual understanding, led us to utilize it for classification.

The architecture of our proposed model is displayed in Fig. 3.2. Our model is an enhancement of “BERTForSequenceClassification”, an implementation of BERT provided by the Hugging Face Transformers library [87], which simplifies the fine-tuning process for sequence classification tasks. We expanded this model to accommodate dyadic interactions by incorporating two inputs: one for parent acts and another for child acts.

Our model utilizes a two-step classification process, consisting of two modules. The first module identifies dyadic acts, specifically the *reflection* class, while the second module recognizes parent acts. In the first module, the model independently processes two input sequences via the RoBERTa model, yielding two embedding vectors, “output\_parent” and “output\_child”. The pooled output from this module is a concatenation of the RoBERTa model’s outputs for the parent and child acts. This concatenated output is then normalized using layer normalization, passed through a dropout layer, and forwarded to a classification

layer to generate class logits. Cross-entropy loss is then calculated between the predicted class probabilities (logits) and the true labels. The second module follows a similar process, but without concatenation of the child acts. Finally, the model concatenates the outputs as logit vectors.

The RoBERTa-based model was fine-tuned over 5 epochs with a learning rate of 1e-5 using the AdamW optimizer. A batch size of 3 was utilized for the fine-tuning process, and cross-entropy was consistently chosen as the loss function.

### 3.5 Evaluation, Results, and Discussion

To evaluate the performance of the models, we used accuracy and weighted F1-measure on the test set. These metrics are defined as follows:

The accuracy is the proportion of correctly classified samples out of the total number of samples in the test set. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the proportion of true positives (TP) to the total number of positive predictions (TP + FP). It represents the accuracy of positive predictions. It is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

The recall is the proportion of true positives (TP) to the total number of actual positive cases (TP + FN). It represents the ability of the model to detect positive cases. It is calculated as:

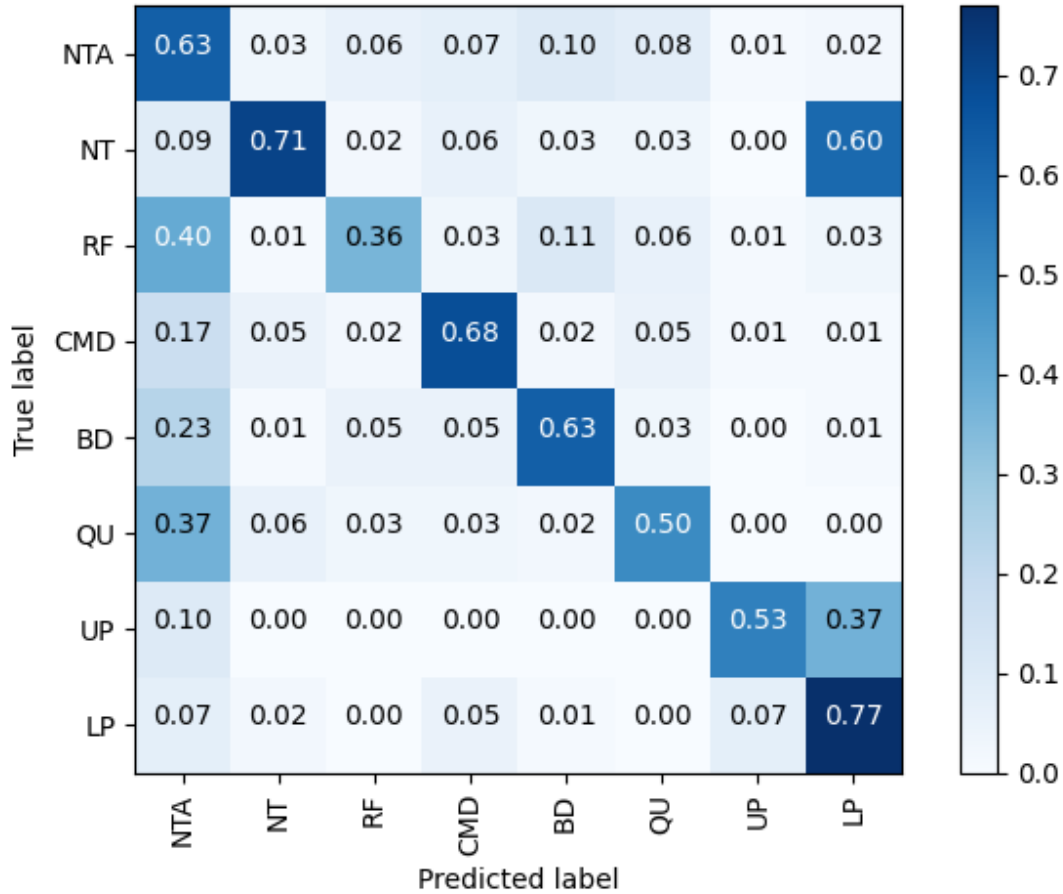


Figure 3.3: Performance results for the ML-based model text

$$Recall = \frac{TP}{TP + FN}$$

The F1 measure is the harmonic mean of precision and recall, which gives an overall measure of the model's performance. It is calculated as:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Fig. 3.3 illustrates the accuracy results for the classical ML-based model, while Fig. 3.4 displays the performance accuracy for one of the better-performing deep learning-based

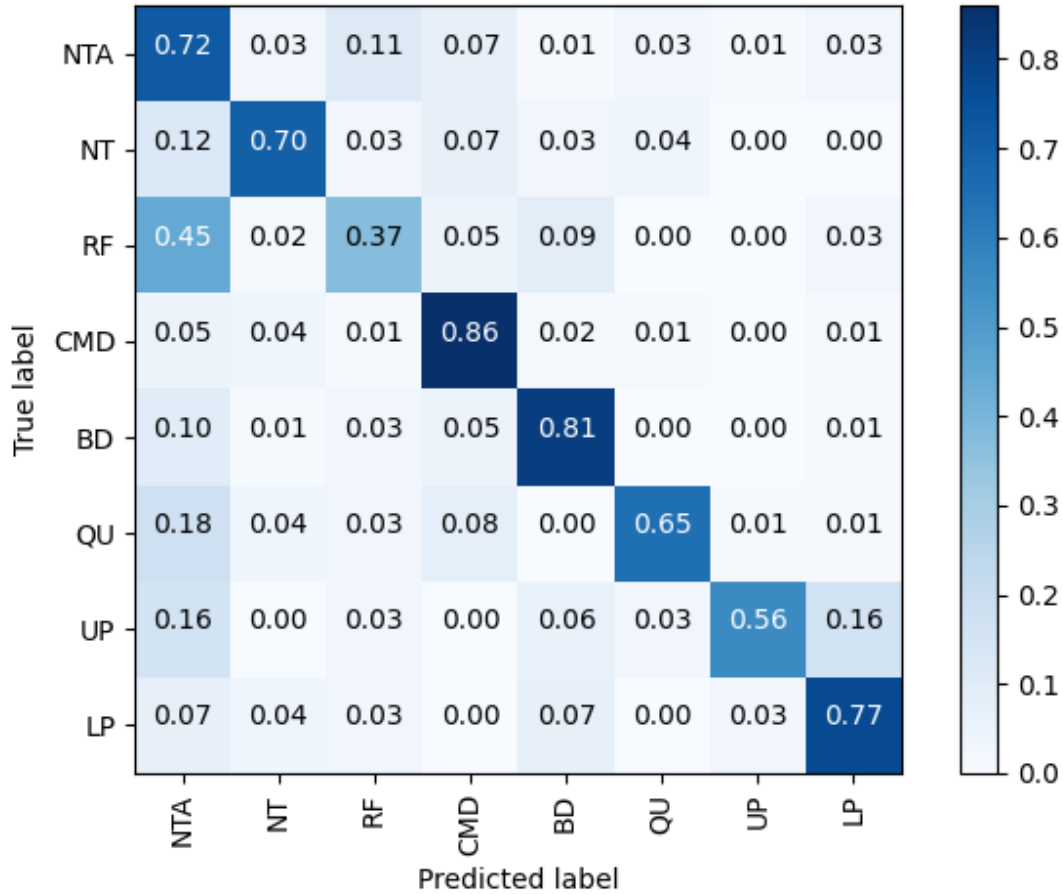


Figure 3.4: Performance results for the deep learning-based model RCNN text [1]

models, the RCNN [1]. As indicated in these matrices, both the deep learning and ML-based models struggled to distinguish between certain classes with intricate similarities that could not be captured solely through the semantics of individual words. This includes classes such as *unlabelled praise (UP)* and *labelled praise (LP)*, or *reflection (RF)* and *neutral talk (NTA)*. Apart from performance, this issue could potentially compromise the overall system quality. For instance, [6] noted that false positive reflections could provide parents with an inaccurately positive impression.

The performance of the proposed model is represented in Fig. 3.5. As anticipated, this

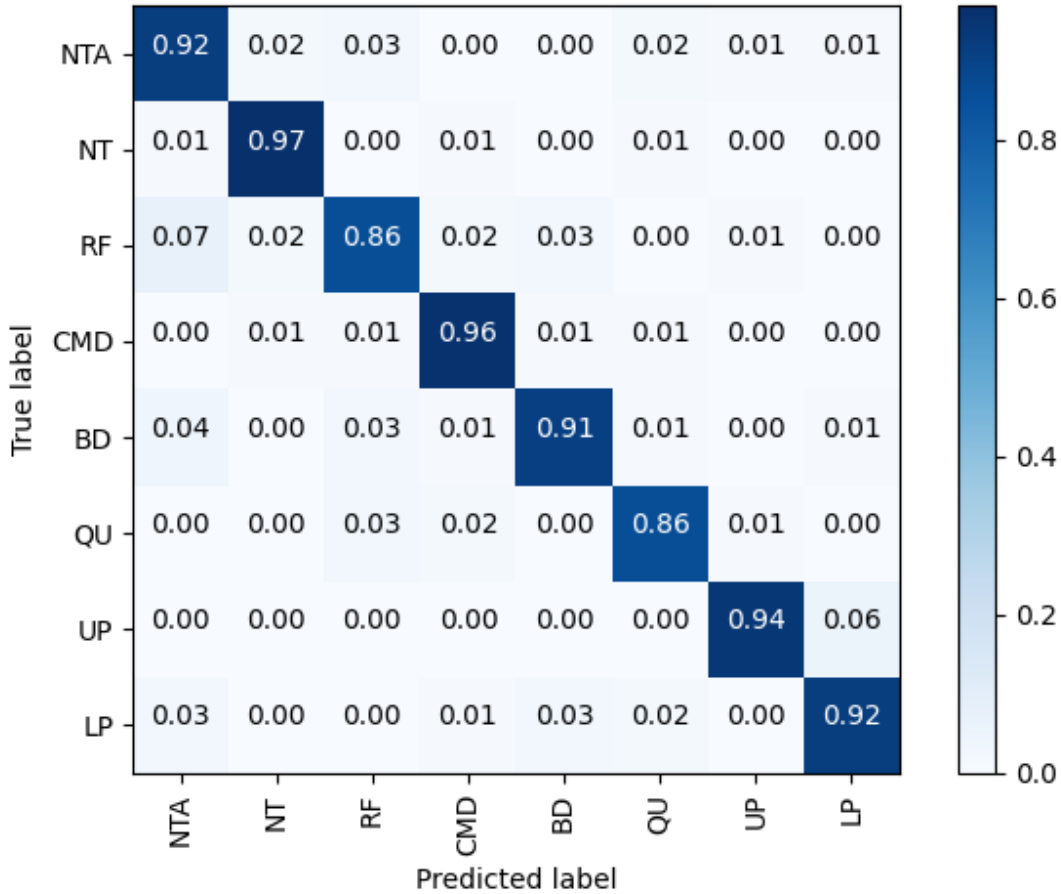


Figure 3.5: Performance results for the proposed model-text

model outperformed others in classifying validation data. The results show that interdependent labels were correctly differentiated. Additionally, the model achieved competitive results for the *question* class without necessitating audio features or additional audio processing. Furthermore, we observed that the model could identify intricate emotional and structural patterns within the text, even without incorporating additional syntactic features or custom features such as LIWC and POS tagging. This suggests that regular feature representations like LIWC, POS tagging, and others may not contribute significant additional knowledge to the model. Various explainability and interoperability solutions could address

this issue.

In a manner similar to [16], we utilized [88] to visualize how the model comprehends the semantic and structural meaning of the text. For instance, [77] discussed in their work on a different text classification task how several heads in their fine-tuned BERT model seemed to focus on the structural context of adjectives and adverbs. Consequently, adding extra grammatical, syntactic, and emotional features based on tokens might not introduce new facets of text meaning to the context.

Table 3.2: Comparison of evaluation performance for the text model

<b>Model</b>	<b>Validation Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-measure</b>
Performance reported in [6]	0.78	0.79	0.77	0.79
ML-based	0.62	0.64	0.61	0.62
RCNN [1]	0.69	0.71	0.68	0.69
CNN [83]	0.68	0.71	0.68	0.69
Proposed Model	<b>0.90</b>	<b>0.92</b>	<b>0.91</b>	<b>0.92</b>

The performance of these models, as demonstrated in table 3.2, is inferior compared to our proposed model. Classical ML models like those presented in [6] suffer from a lack of sufficient labelled data, particularly in the case of rare n-grams. Therefore, they require more advanced features to handle the intricacies of a complex domain.

Meanwhile, deep learning-based models in this field generally rely on a combination of concepts from CNNs and RNNs. The CNN identifies the most significant n-grams, while the

RNN calculates a weighted combination of all the words in the text. Thus, a blend of the two can capture both semantic and structural aspects within the text. However, in certain instances, this approach still falls short of delivering superior results compared to ML-based models. This is partly because word embedding models employed in deep learning-based solutions demand a larger amount of labelled data to adapt to the domain. Additionally, unlike BERT-based models, these models are context-independent and do not consider the word’s position in a sentence.

Regarding interpretability, Fig. 3.6 presents a local attribution sample from the proposed model. The examples shown are true positives (correctly identified as a PCIT category); terms highlighted in green contribute to a positive prediction, while red ones detract from it. These attributions demonstrate that the proposed model’s domain adaptation and fine-tuning were able to discern specific structural, semantic, and grammatical factors related to each class, obviating the need for additional POS tagging and LIWC features. Hence, there is no need to add customized features to an ML-based approach or to rely on CNNs and RNNs to capture emotional and structural meaning.

For instance, in the *negative talk* category, phrases containing negative structures (e.g., “not correct”) and absolute expressions (e.g., “always”) are accentuated. This example underscores the importance of grammar and structure to the target class. In the *command* category, words denoting a specific object (e.g., “toys”) or time (e.g., “now”) in a given context are crucial to the target class. In the *question* category, the model could discern the significance of starting a sentence with a verb, though it is not a definitive condition. And in the *labelled praise* category, the model could understand when positive emotional states (e.g., “perfectly”) are tied to a specific topic (e.g., “that”), which differentiates it from

negative_talk	[CLS] no it is n ' t correct [SEP]
negative_talk	[CLS] you are always knocking things over [SEP]
negative_talk	[CLS] quit getting out of your chair [SEP]
command	[CLS] i need you to pick up the toys [SEP]
command	[CLS] why do n ' t we build castles now [SEP]
command	[CLS] you have to put the cars away now [SEP]
behavior_description	[CLS] you are jumping [SEP]
behavior_description	[CLS] you seem happy that you fixed it [SEP]
behavior_description	[CLS] you are not listening [SEP]
question	[CLS] would this help [SEP]
question	[CLS] have you ever seen a monster truck [SEP]
question	[CLS] even the rocket [SEP]
unlabelled_praise	[CLS] so creative [SEP]
unlabelled_praise	[CLS] good job [SEP]
unlabelled_praise	[CLS] wow cool [SEP]
labelled_praise	[CLS] you did that perfectly [SEP]
labelled_praise	[CLS] thank you for playing nicely [SEP]
labelled_praise	[CLS] great job respecting the toys [SEP]

Figure 3.6: Explanations for the proposed text model on a true positive sample set

*unlabelled praise.*

### 3.6 Conclusions

PCIT, or Parent-Child Interaction Therapy, is a therapeutic approach that aims to enhance parent-child relationships by instructing parents on how to engage more effectively with their children for the betterment of their psychological and behavioural development. Our research underscores the potential of AI in streamlining PCIT evaluations, addressing



the time and resource limitations often encountered in these assessments.

We used the “SpecialTime” system dataset, which includes over 6,000 parent-dialogue acts, annotated according to the DPICS. With the assistance of the latest developments in Natural Language Processing, we achieved a validation accuracy of 90%, significantly surpassing previous outcomes.

Due to the intricate and ambiguous characteristics of the classes, conventional ML techniques proved to be labour-intensive. As such, we shifted towards transfer learning, leveraging a fine-tuned Transformer-based model that surpassed other methods. Furthermore, our findings revealed that our approach can deliver high performance even in the absence of vocal features.

Interestingly, we found that child acts had minimal influence on the classification outcome, with the exception of a single label. This effectively reduces the added complexity that dyadic interactions typically introduce. We managed to accomplish this by proposing a two-step classification model that utilizes child acts only for the identification of dyadic labels, while parent acts are employed to discern other labels.

### 3.7 Acknowledgement

We would like to express our sincere gratitude to Professor Monica Oxford from the Department of Child, Family, and Population Health Nursing at the University of Washington and Professor Nicole Letourneau from Nursing and Cumming School of Medicine (Pediatrics, Psychiatry, and Community Health Sciences) at the University of Calgary for their invaluable guidance, feedback, and support throughout the course of this research.

# Chapter 4

## A Model for Parent-Child Interactions Analysis from Text and Audio

### 4.1 Abstract

The quality of parent-child interactions is a critical foundation for future social and emotional development and well-being. Among tools that assess the quality of these interactions, the Parent-Child Interaction Teaching Scale (PCITS) is a well-established and effective tool. Identifying parent-child behaviors early is a significant challenge in utilizing PCITS, to help parents address initial behavioral issues. However, the necessity for resource-intensive and time-consuming manual evaluations limits the accessibility of these assessments.

In our research, we delved into the complexities and challenges of automating this process using AI for audio and text modalities. We proposed a solution to categorize the primary behavior types in PCITS. We believe that our study is the first to introduce a model that uses a novel framework to incorporate multiple behavioral factors, allowing us to analyze emotions and psychological elements in parent-child interactions through both audio and text modalities.

Our proposed model can discern and detect the audio's semantic features and the linguistic characteristics, including both semantic and syntactic aspects, in parent-child interactions according to the PCITS scale. The model uses an ensemble learning approach that integrates various aspects of recognition within this scale. Compared to similar efforts, our evaluation

results showed an improvement in performance when utilizing this approach.

## 4.2 Introduction

The Parent-Child Interaction Teaching Scale (PCITS) is a well-established and trusted tool for evaluating the quality of interactions between parents and their children. It primarily assesses how caregivers communicate with their children, allowing therapists to identify anomalies and guide parents in improving their interactive behaviors. However, the significant time and effort required for these interventions is a major limitation. Consequently, the central research question we pose is how AI could be leveraged to automate some of these evaluations, by assessing interaction quality based on linguistic and audio features in dynamic dialogues.

Current research on human behavior analysis through text and audio mostly focuses on primary categories such as basic sentiment and emotions, using publicly available data. Despite the crucial societal benefits of PCITS, there is a noticeable lack of studies and data in this area. This gap is partly due to ethical considerations around privacy and partly due to the intricacies of data collection that require expert qualification and training to gain the needed knowledge.

In the context of the verbal labels in the PCITS scale, we have compiled a dataset assembled by skilled coders. The labels, which include *imperative instruction*, *explanatory instruction*, *broad praise*, *task-related praise*, *cheerleading*, and *negative comment*, rely heavily on both semantic and syntactic features in language and audio analysis. For instance, *imperative instruction* generally starts with a command or verb, while *explanatory instruc-*

*tion* provides more task-specific descriptions but has a similar structure. *Task-related praise* is a positive verbal action targeting a specific activity, making the positive sentiment context-dependent, while *broad praise* is a more general positive statement. *Cheerleading* encourages the child in a unique way but shares a similar sentiment with *broad praise*. The *negative comment* label may involve the use of a negative word or vocal tone. The specifics of these labels are detailed in Table 4.1.

Given these definitions and examples, we have examined these labels from two semantic and syntactic perspectives, focusing on the dual modalities of language and audio. We then proposed a multi-modal approach to combine different modalities and viewpoints.

Table 4.1: PCITS text dataset examples

<b>Class</b>	<b>Description</b>	<b>Example</b>
Imperative Instruction (II)	A given instruction or command	(1) Catch it. (2) Put it here. (3) Do this.
Explanatory Instruction (EI)	A given instruction to explain why something needs to be done a certain way	(1) I want to see if you can turn the page. (2) Do you want to try it now? (3) I want to you move the car by pulling the string.
Continued on next page		

Table 4.1 – continued from previous page

Class	Description	Example
Broad Praise (BP)	Indicates general praise given	(1) Good job. (2) You're such a good worker. (3) I'm proud of how hard you worked today.
Task-Related Praise (TP)	Indicates praise given by the adult to the child specifically related to the task	(1) You figured that out fast. (2) Good job. Did it. (3) There you go.
Cheerleading (CH)	Indicates a statement of encouragement or motivation	(1) You can do it too. (2) Keep trying. (3) Just one more.
Negative Comment (NC)	Indicates a negative comment or feedback	(1) No, you can't eat the frog. (2) Open your hand and Oh nope. (3) Don't eat it.

#### 4.2.1 Semantic characteristics

Semantics pertains to the meaningful context of something, with the semantics of dialogue often depending on emotion or sentiment in relation to that context. Sentiment analysis,

as discussed by [89], involves the examination of feelings, attitudes, evaluations, emotions, and thoughts regarding various entities such as products, services, companies, individuals, issues, events, and their characteristics. [90] classified emotions into six basic categories: anger, disgust, fear, joy, sadness, and surprise.

There are multiple evolutionary, neurological, and psychological approaches to analyzing human emotions, aimed at deriving complex labels from these basic ones. For instance, [91] proposed a theory that subdivides emotions into 24 categories, some comprising two emotions and others composed of three primary emotions. However, this approach involves a complex task in feature engineering and necessitates a solution for recognizing and mitigating the correlation between labels. Furthermore, in our case, it is impossible to apply traditional machine learning algorithms to identify a relationship between domain-specific labels and basic or universal classes due to the complexity of the label definitions.

A viable alternative is to leverage deep learning and transfer learning, which can reduce the efforts needed for feature engineering and domain-specific rules. CNN models have recently shown remarkable success in identifying emotional states in both text and audio, as cited in [13]. The key to this success lies in the nature of CNNs that slide through n-gram features, selecting the most discriminative language fragments in max-pooling that convey the most emotion in the text.

Regarding audio semantic characteristics, vocal features like log-scale Mel spectrograms are frequently utilized in emotion and semantic recognition tasks. These spectrograms mirror how humans perceive sound and are widely employed in speech and audio processing applications. The use of a visual representation of vocal features makes CNNs a suitable choice for analyzing audio, as mentioned in [92]. For example, [14] utilized pre-trained models like

ResNet for audio semantic classification.

#### 4.2.2 Syntactic characteristics

Syntactic attributes refer to the style, syntax, and grammar of speech, forming a key part of the NLP task in this study. Within this domain, the sentence type and structure are crucial for classifying certain categories. For instance, the sentence “Who do you think you are talking to?” is identified as a type of *negative comment* behavior within parent-child interaction, partially because it begins with the word “who”. As mentioned in [15], sentential or syntactic features such as part-of-speech (POS) tagging are particularly important for these classes. Recognizing patterns like *broad praise* often being found in shorter sentences, and *imperative instruction* commonly appearing in sentences that begin with a verb, can enhance classification effectiveness.

In terms of deep-learning-based approaches, RNNs and their extensions are growing increasingly relevant, given their capacity to capture recurring and temporal linguistic features that represent text structure and grammar [73]. Moreover, Transformer models like BERT have shown their proficiency in detecting syntactic structures [16]. We utilized both options in our research.

In our proposed solution, we combined the two key facets of the problem, i.e., the *semantic* and *syntactic* characteristics, in both audio and text modalities. The challenge thus lies in developing a comprehensive model that encapsulates all these aspects and modalities within a single architectural framework. We designed a meta-model that trains over the outputs of independent modules, each responsible for a specific aspect within a modality, and developed a method to map these to the final result.

The remainder of this chapter is organized as follows: Section 4.3 presents the works surveyed for this study. Section 4.4 explains our model’s architecture and settings, along with some additional features and enhancements. Subsequently, in Section 4.5, we evaluate the individual models in contrast with the ensemble model and discuss our observations. We finally conclude our study in Section 4.6.

### 4.3 Related Work

Among empirical studies in PCITS, systems such as TalkBetter [66] and TalkLIME [67] monitor dialogues and alert when they detect harmful language patterns employed by a parent. In the realm of AI, there exist two primary perspectives: semantic and syntactic characteristics. For each perspective, two broad categories of models are present: machine-learning-based and deep-learning-based.

Numerous machine-learning-based solutions in text modality, such as [6], rely on pure lexicons, and they hold the advantage of being easier to handle, particularly when compared with intricate text representations. The unit of analysis is words, the representation of text is TFIDF, and the most commonly employed features include bag-of-words (BOW), linguistic inquiry and word count (LIWC), and POS tagging. LIWC may be beneficial for semantic or emotional features, while POS-tagging can be advantageous for syntactic features. LIWC is a tool that operates using a dictionary-based methodology, classifying words into predefined groups relevant to syntax and semantics, and subsequently quantifying their frequency. The features offered by LIWC are recognized as useful in detecting behavioral states, such as in depression diagnosis [61]. The model employs a support vector machine (SVM) with LIWC



features, selected based on empirical studies. In another example, [69] utilized LIWC to substantiate their assertion that speakers feeling positive emotions tend to use a higher quantity of positive affect phrases, a more diverse vocabulary, and fewer first-person pronouns.

In parent-child interaction studies focusing on handcrafted features, [71, 72] have applied the breadth of a parent’s vocabulary, linked with a child’s language skills, to identify unique forms of grammatical and emotional word associations. [64] proposed a system designed to educate parents on providing appropriate linguistic stimuli to their children, thus reducing the possibility of language development delays. These systems assist in improving children’s linguistic development by providing parents with insights regarding aspects such as speech volume, proper responses, and lexical diversity. These features are developed based on empirical research.

The effectiveness of machine-learning-based solutions often hinges heavily on extensive efforts in feature engineering. These encoding strategies, however, do not consider the positioning of words within the text or recognize significant links between categories, which can hinder the model’s capacity to generalize [60]. Also, these models can suffer from an insufficient amount of labeled data. While lexicon-based models and one-hot vectors provide clearer and more interpretable results compared to end-to-end deep learning techniques, they can struggle with complex sentences [60] and evaluating the likelihood of uncommon n-grams [13]. Ultimately, traditional models might not fully incorporate the context of the text, including the sequence of words, the emotion expressed, and the meaning and context of parental interactions.

In response to these limitations, advancements in deep learning have prompted the use of variants of CNNs [93] and RNNs [73] to tackle these challenges. Issues are further addressed

by word embedding representation models like BERT [94]. These models, which are trained on vast unlabeled datasets available online, demonstrate the property of semantically similar words having close proximity in the corresponding vector space.

New representation models encounter difficulties when identifying certain types of grammatical and emotional relationships between words [60]. In the context of PCITS, recognizing the syntactic characteristics of the dialogue is crucial. However, for Transformer-based embedding models like BERT, researchers have proposed solutions that include fine-tuning these models to better recognize structural context within text [16]. Another challenge with embedding models lies in sentiment and emotion analysis as they typically only model the context of words, overlooking the sentiment information.

This problem can be mitigated by capturing the word’s position in a sentence to better resolve context, a strategy employed by BERT [60]. Another method is to merge features from different perspectives. For instance, [95] has combined a lexicon-based model with document embedding features in an effort to detect symptoms of mental illnesses from interview text. Similarly, the RCNN model proposed by [1] combines RNN with CNN, where the convolutional layer operates recurrently to capture the most significant local features in terms of temporal characteristics.

The approach of blending part-of-speech features with text embedding within an attention-based CNN model is utilized in [96, 97] to detect behavioral status within text. In such models, the convolutional layer is built upon the word vector representation obtained from an unsupervised neural language model. RNNs and their variations maintain relationships between words in a sentence by leveraging historical data. For instance, RNNs tend to perform better when text contains negating phrases such as “won’t” and “miss” [73]. As

pointed out in [98], while CNNs extract the most meaningful n-grams, RNNs compute a weighted combination of all words in the text. By integrating these models, a more accurate representation is possible, as exemplified by BERT and its variants [74].

When considering acoustic semantic features in parent-child interaction studies, numerous works utilize vocal features like loudness, jitter, and energy in the parent’s voice [62]. Similarly, in studies on depression detection, features are learned and incorporated into textual features [63]. In [6], the parent’s voice tone is used to detect questions by calculating the initial derivative of the pitch curve from the final half second of the speech segment. Apart from empirical approaches, the use of CNNs for analysis is suggested in various other fields [14,99,100]. Although CNNs were originally developed for image processing, the visual spectrogram of vocal features allows for effective classification tasks through CNN advancements. Some researchers even utilize transfer learning to leverage pre-trained models such as ResNet variants for audio analysis [14].

## 4.4 Methodology

### 4.4.1 Dataset

The video content was curated and annotated by trained research assistants utilizing a web-based labeling interface that was developed and launched during the early stages of this study. Appendix C provides information about the labelling system. This tool proves beneficial for both dataset preparation and its automation. Subsequently, the videos were converted to audio and transcribed using the Amazon Transcribe service<sup>1</sup>.

---

<sup>1</sup><https://aws.amazon.com/transcribe/>

Unlike the approach used in [6], our transcriptions include punctuation marks like “question marks,” which aid in the development of more refined structural features within the text. The presence of punctuation marks also simplifies the sentence segmentation process, maintaining the models’ granularity based on sentences without needing additional audio or text processing.

We discovered that only a single caregiver was involved during the recorded interaction, and any vocal sounds made by the child were disregarded as noise by the transcription service. Consequently, the caregiver’s speech became the sole significant component, making it feasible to transcribe text from a single participant (the parent). This presumption assisted in eliminating dyadic complexities from the text analysis in PCITS.

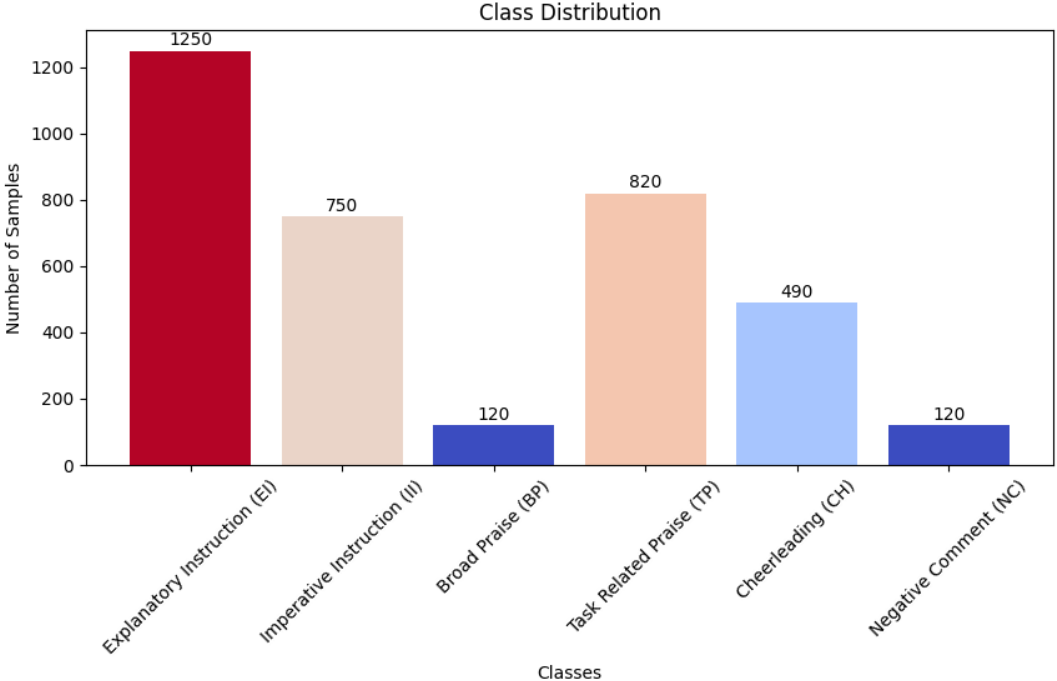


Figure 4.1: Dataset class distribution - audio/text

Fig. 4.1 illustrates the quantity of samples within our dataset. Two labels, *broad praise*

and *negative comment*, were found to have less representation. To augment these two categories, we utilized back-translation [101], generating approximately 600 additional samples.

We also employed ChatGPT <sup>2</sup> for augmentation, providing it with precise definitions of the labels, their distinctions from other labels, and a series of samples. This allowed the generation of 600 similar instances, following the concept of few-shot prompt engineering [102].

Afterward, we filtered out irrelevant and similar augmented samples, reducing the final count for these two labels to approximately 500. To eliminate semi-duplicate samples, we utilized Sentence-BERT [103] to create embedding vectors of samples. Using cosine similarity as a metric on these vectors facilitated the removal of the most similar samples.

#### 4.4.2 Audio classification

According to our experiments, the audio classification technique we used was unable to differentiate between the labels “broad praise”, “task-related praise”, and “cheerleading”. This could be due to the fact that these labels are distinguished by the syntactic structure within the text. Therefore, we decided to group these three labels into a single general label, which we named “positiveness”.

To extract features from audio samples, we utilized a log-scale Mel spectrogram with 64 Mel frequency bins. We applied a hop length of 0.01 seconds and a window length of 0.02 seconds between two consecutive Hanning windows. These features were then inputted to a pre-trained ResNet18 [104] model on the ImageNet dataset. The model was adapted by modifying the first layer to accept grayscale images and replacing the last fully connected

---

<sup>2</sup><https://chat.openai.com/chat>

layer with a dropout regularization of 0.25 probability and a linear layer with output features equal to the number of classes. The original model contained 11,689,512 parameters, while the updated model contained 11,172,356 parameters. We trained the model for 50 epochs using a batch size of 8 and the Adam optimizer with a learning rate of 0.0001 and momentum of 0.9.

We used the cross-entropy loss as the loss function for all classifier modules. To account for class imbalance, we adjusted the loss function with a balanced weight factor, which was calculated based on the distribution of samples. The calculation was done using the method proposed in [105] in PyTorch.

For each model, we employed a grid-search cross-validation strategy across 5 folds to fine-tune the hyperparameters. The chosen hyperparameters were those that performed best in terms of accuracy on the validation set.

#### 4.4.3 Text classification

The proposed model employs three distinct methods for text classification:

**Deep-learning-based model** We followed the methodology proposed in [13] to perform text classification using a CNN+RNN architecture. Firstly, we used a recurrent neural network to analyze the POS tagging features of the text. This approach was chosen due to the sequential and structural nature of the POS tags, which can help in capturing the sequential characteristics of the text. The tags we focused on included “DT (Determiner)”, “JJ (Adjective)”, and “MD (Modal)”. Secondly, we utilized problem-specific features, including 9 syntactic features derived from the POS tags. These features captured some special states in the domain that can affect the final classification step, such as detecting “imperative instruc-

tion” and “explanatory instruction” categories. Thirdly, we employed a convolutional neural network to process word embedding features that capture the most significant semantic features of the text. The objective of this step was to capture the most semantically relevant parts of the text. Finally, we concatenated the output of the convolutional network with that of the recurrent network and the classifier network and then used a fully connected layer to classify the final output. To create sentences, we selected certain tags from POS tagging by NLTK <sup>3</sup>, including “PRP”, “NN”, and “MD”. As these tags are sequential, we employed an LSTM network with a dropout layer to avoid bias and hidden unit co-adaptation [93]. We then used a fully connected layer with RELU activation to choose the most significant sentence composition features. To create word embeddings, we used the word2vec algorithm trained on GoogleNews [106]. This algorithm captures the contextual meaning of words by considering their co-occurrence with other words in a given corpus. We utilized a CNN model with two convolutional layers and max-pooling followed by a dropout layer and a fully connected layer activated by the RELU function to extract the most important contextual features. These features were then classified by another fully connected layer. The CNN was trained for 20 epochs using the Adam optimizer with a learning rate of 0.001.

**Fine-tuning** We utilized a BERT model as the second approach for text classification, aiming to fine-tune it for our specific task. The rationale behind using pre-trained models like BERT is that they can perform better in situations where data is limited, compared to training from scratch. We employed the “BERTForSequenceClassification” implementation from the Transformers library [87] provided by Hugging Face. The optimizer used was AdamW [85], which is a variation of Adam that uses decaying weight. The model was fine-

---

<sup>3</sup><https://www.nltk.org/>

tuned for 30 epochs with a batch size of 8. The number of parameters in the fine-tuned model was 109,486,854.

**Transfer-learning** A dataset was introduced in [6] containing approximately 4000 records with labels such as “negative talk”, “question”, “command”, “behaviour description”, “un-labeled praise”, and “labelled praise”. We conducted research on this dataset to create a model that would achieve higher accuracy than the one reported in the original work [107]. We then used this model as a feature generator module. To find a non-linear relationship between the output features of the tuned model and the final target, we developed a neural network. We decided to exclude “cheerleading” from the dataset for the second classifier, as it was not significant for this label, possibly because there was no similar label in the first dataset. The neural network consisted of two fully connected layers with a ReLU and dropout layer between them. The first layer had dimensions of  $8 \times 64$  and the second layer had dimensions of  $64 \times 6$ . The number of epochs for this classifier was 20, and we used the Adam optimizer with a learning rate of 0.001.

#### 4.4.4 Combined model

Fig. 4.2 shows the overall architecture of the model. The results of the four audio and text classification modules provided different perspectives and types of information for the classification task. To combine these outputs, we used an ensemble learning model. Ensemble methods aim to create a model that reduces the biases of a single machine-learning algorithm [108]. Among different ensemble techniques, we opted for a stacking approach, which involves using a regression model to learn the weights for a weighted averaging model.

Thus, we denoted  $p_{i,j}$  as the likelihood of model  $M_i$  for assigning class  $j$  to a given



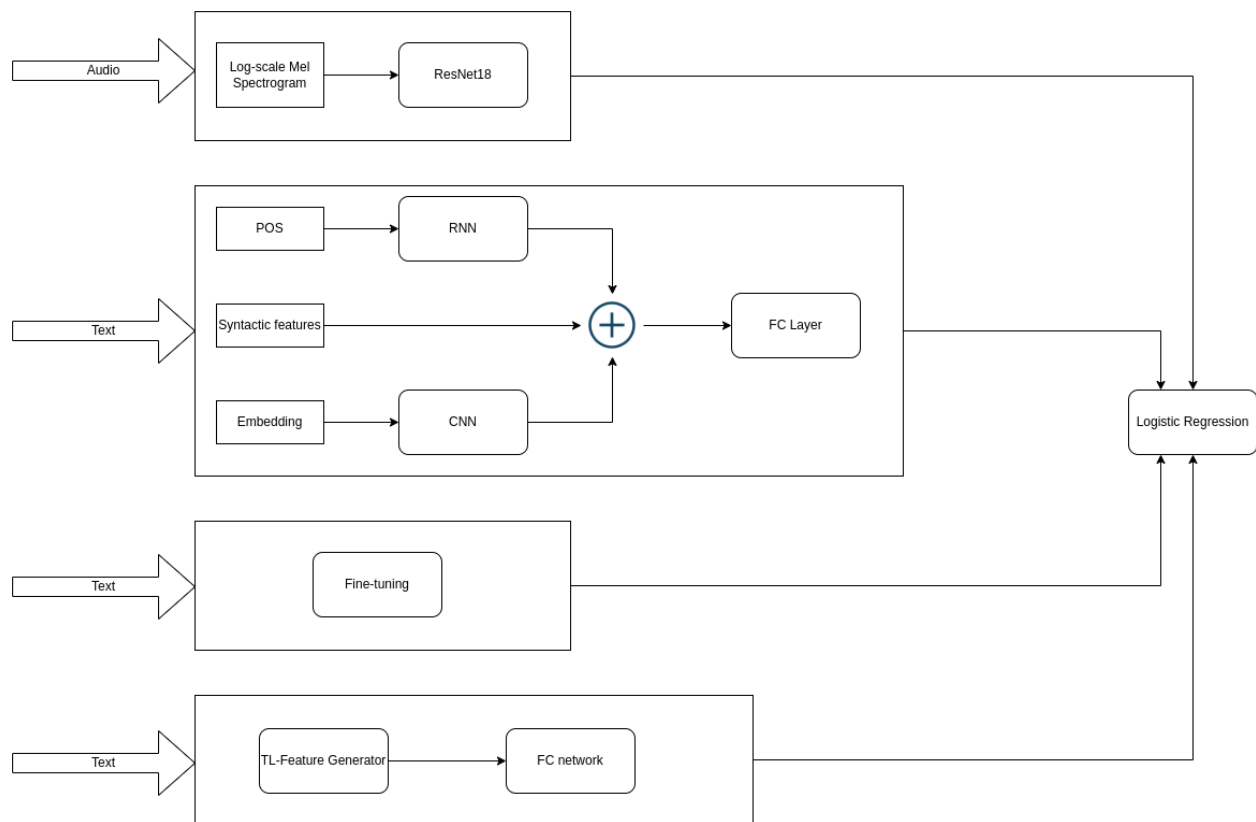


Figure 4.2: Model architecture for text/audio modalities

input, where  $i \in \{1, 2, 3, 4\}$  and  $j \in \{1, 2, \dots, K\}$ , where  $K$  is the total number of classes (six). The value of  $p_{i,j}$  was zero if the class  $j$  was not among the target labels of model  $M_i$ . For example, the target “cheerleading” was ignored for the transfer-learning-based model. Also, for the audio classification model, all outputs corresponding to the three labels “broad praise”, “task-related praise”, and “cheerleading” were assigned the same output value for the label “positiveness”.

To combine the output of the four modules, we utilized an ensemble learning model. The ensemble model was designed to learn the relationship between  $M_i$  and the final output  $y$  using the following equation:

$$\hat{y} = \text{softmax} \left( \sum_{i=1}^4 w_i \cdot \text{softmax}(M_i(x)) \right) \quad (4.1)$$

Here,  $\hat{y}$  represents the predicted probabilities of the ensemble model for a given input  $x$ , and  $\text{softmax}(M_i(x))$  denotes the vector of probabilities obtained by passing the logits through a softmax function, resulting in class probabilities. The weights  $w_i$  were learned by training a LogisticRegression model on the concatenated outputs of the four models for each input, as explained above.

The softmax function is defined as:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (4.2)$$

where  $z_i$  is the  $i$ -th element of the model’s output vector, and  $K$  is the number of classes. The softmax function transforms the model’s outputs into positive values that sum to one, representing class probabilities. We applied softmax to the output probabilities of each

classification module to obtain a normalized vector.

## 4.5 Evaluation and Results

To build the classification models and the ensemble model, we split the dataset into two parts. The first part, which comprised 80% of the data, was further divided into training, testing, and validation subsets in a 60%, 20%, and 20% ratio, respectively. This part was used to build the four classification modules. The second part, which contained 20% of the data, was used to build the ensemble model. This part was divided into training and validating subsets in an 80% to 20% ratio. We used k-fold cross-validation with 5 folds to train the ensemble model. However, we did not include augmented audio samples in the portion of the dataset assigned to the ensemble model because there were no corresponding audio samples for the augmented text samples.

Table 4.2: Text-audio performance results

Label	4.4.2 <sup>1</sup>		4.4.3 <sup>2</sup>		4.4.3 <sup>3</sup>		4.4.3 <sup>4</sup>		4.4.4 <sup>5</sup>	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
<b>Imperative Instruction</b>	0.43	0.59	0.60	0.65	0.29	0.45	0.53	0.69	<b>0.80</b>	<b>0.82</b>
<b>Explanatory Instruction</b>	0.61	0.76	0.65	0.71	0.69	0.82	0.68	0.81	<b>0.80</b>	<b>0.82</b>
<b>Broad Praise</b>	0.72	0.84	0.30	0.43	0.07	0.13	0.58	0.73	<b>0.65</b>	<b>0.77</b>
<b>Task-Related Praise</b>	-	-	0.45	0.57	0.59	0.74	0.63	0.77	<b>0.78</b>	<b>0.81</b>
<b>Cheerleading</b>	-	-	0.38	0.50	-	-	0.46	0.63	<b>0.55</b>	<b>0.62</b>
<b>Negative Comment</b>	0.15	0.20	0.50	0.54	0.68	0.81	0.67	0.80	<b>0.80</b>	<b>0.82</b>

<sup>1</sup> Audio classification model, <sup>2</sup> Deep-learning-based model, <sup>3</sup> Transfer-learning-based model, <sup>4</sup> Fine-tuning model, <sup>5</sup> Ensemble model

Table 4.2 displays the results of five models in terms of their accuracy and F1 scores for various labels.

The models include four different classification modules: deep-learning-based, transfer-learning-based, fine-tuning, and audio classification, as well as the ensemble model. The accuracy and F1 scores for each label are presented, with the best results for each label highlighted in bold.

Table 4.3 presents the overall performance metrics, which include validation accuracy, precision, recall, and F1 score, calculated in a weighted manner for all models.

Table 4.3: Overall evaluation performance between models in text-audio modalities

Model	Validation Accuracy	Precision (weighted avg)	Recall (weighted avg)	F1-measure (weighted avg)
Audio classification	0.62	0.60	0.61	0.60
Deep-learning-based text classification	0.54	0.64	0.62	0.62
Transfer-learning-based text classification	0.54	0.60	0.48	0.49
Fine-tuning text classification	0.60	0.70	0.61	0.63
Ensemble model	<b>0.77</b>	<b>0.83</b>	<b>0.78</b>	<b>0.79</b>

According to the results, the ensemble model showed better performance in both accuracy and F1 score compared to the other models. The audio classification module demonstrated better performance on the three labels related to “positivity”, including “broad praise”, “task-related praise”, and “cheerleading”. Although the model could not differentiate between these three labels, we considered the same output for all three in this module. The deep-learning-based model showed better performance on the “imperative instruction”, and “explanatory instruction” labels, which may be attributed to capturing the structural context that is more important for these labels. The transfer-learning-based model demonstrated better performance on the “explanatory instruction”, “task-related praise”, and “negative comment” labels, which may be attributed to having labels with similar definitions in the dataset used for this module. The fine-tuning model demonstrated more consistent performance across all labels, although it did not achieve the highest accuracy.

In summary, the ensemble model demonstrated superior performance compared to other models, achieving the highest accuracy and F1 scores on all six labels. The advantage of using an ensemble model was the ability to incorporate multiple aspects and modalities of the classification task, including audio and text, semantic and structural context. Additionally, the correlation between labels posed a challenge for a single classification model to predict a single target. By using a normalized vector of output probabilities for each label, the ensemble model could leverage the complex relationships between labels to improve performance.

## 4.6 Conclusions

The aim of this study was to examine the verbal behavior of parents during interactions with their children, using the PCITS measure as a reference [17, 109]. In the early stages of the research, we developed a labeling system. Given the complexities of the domain, categories are dependent on both syntactic and semantic characteristics from text and audio modalities. Therefore, we implemented four different classification modules to be merged in a meta-model.

We utilized a combined CNN+RNN model with additional structural features to focus on domain-specific feature analysis based on semantic and syntactic context. Another model was used to conduct classification on a similar dataset and transfer the learned knowledge. Furthermore, we fine-tuned another model on the

text dataset, leveraging recent advancements in the Transformers architecture.

For the audio modality, we employed a model pre-trained on image classification tasks and trained it on the vocal spectrogram of the audio samples. Finally, we integrated the four classification modules into a logistic regression classification model to amalgamate the outputs of all models.

By adopting an ensemble classification approach for training, the model demonstrated satisfactory performance. To summarize, our research focused on evaluating parent-child interaction using the PCITS scale, with a focus on two primary modalities - audio and text, and two types of features - semantic and syntactic. We proposed a model that is capable of incorporating all these modalities and aspects, and the model has demonstrated comparable performance to similar studies in this field.

## 4.7 Acknowledgement

We would like to acknowledge the generous support of many organizations that funded the EQUIP project. We thank AMS Health Care for the Fellowship in Artificial Intelligence and Compassion, UCalgary for the VPR Catalyst Grant, Alberta Children’s Hospital Research Institute for the Catalyst Seedling Award, and the Social Sciences and Humanities Research Council (SSHRC) for the Insight Development Grant. These sources supported the preparation, labelling, and analysis of video data codes. We also express our gratitude to CIHR for the source PCITS videos (University of Calgary REB# 16-1811). We extend our heartfelt thanks to the CIHR study participants that make this work possible. We also thank Lyndsay MacKay, Jason Novick, Jennifer Black, Alexa Toews, Chris Street, Tian Westland, Linnea Davison, Harleen Sanghera, and Carl Dizon.

# Chapter 5

## Conclusion and future works

### 5.1 Summary and Conclusion

The research provided different AI methods for analyzing parent-child interaction quality. The analysis included three modalities: video, audio, and language. It also investigated various aspects of each modality.

Chapter 2 relates to the analyses of parent-child interactions in view of the video modality. This could be considered a part of human behaviour analysis (HBA) problems in AI. We tackled this HBA problem using two methods: data-driven and knowledge-based. Regarding the literature on HBA problems, there are several challenges associated with analyzing such interactions in 2D video files, such as occlusion and camera shake. To address these challenges, a tracking algorithm was developed to detect the skeleton joints of both parent and child actors. Additionally, semantic-based features were introduced to handle limited labelled data and occlusion. The features were refined through feature selection, which decomposed complex behaviour into simple activities occurring in body parts. This decomposition enabled the model to be trained in reverse by recognizing actions from semantics in body parts. The performance of the proposed model was evaluated using transfer learning through a data-driven approach and a statistics-based model, consistent with previous research.

Chapter 3 relates to the analyses of parent-child interactions in view of text modality in a similar dataset. The purpose of the chapter was to improve parent-child interactions by implementing a treatment called PCIT that teaches parents how to communicate more effectively with their children for improved mental and behavioural development. In order to accomplish this, a method was developed to classify conversational texts between parents and children into categories that encourage high-quality interactions using the "SpecialTime" system dataset. Although the claimed accuracy of the PCIT classification task on this dataset was 79%, this research resulted in an overall accuracy increase of 11%. Surprisingly, good performance was achieved without the incorporation of any vocal features, and it was discovered that the

child's actions had minimal effect on the classification outcome, with one category being an exception. This means that similar to the case in this research, dyadic features are not required in language modality. This study suggests that transfer learning from a BERT model that has been fine-tuned produces superior results compared to using either deep learning or ML techniques on this dataset. The trained model was transferred to the work in research in chapter 4.

Chapter 4 relates to the analyses of parent-child interactions in view of text and audio modalities. The aim of this chapter was to investigate the language used by parents when interacting with their children using the PCITS measure. To achieve this, both audio and text data in view of semantic and syntactic features were used. Four different classification models were utilized in this chapter: a CNN+RNN model that focused on the semantic and syntactic context of the text, a transfer-learning model that incorporated knowledge from the trained model into the text in chapter 3, a fine-tuning model that used the latest advances in transformer architecture in text classification, and a pre-trained model that was originally intended for classifying images but was retrained to classify vocal spectrograms from audio samples. The outputs of these models were combined using logistic regression to create an ensemble approach, which showed satisfactory performance and will be integrated into a larger framework for analyzing parent-child interactions.

## 5.2 Limitations

The main limitation of this study is the lack of labelled data. Most of the experiments in this research were done in the last eight months when a team of six coders started to provide these data. As discussed for the video modality in chapter 2, the complexities and higher dimensions of this modality require probing into knowledge-based solutions such as hand-made features and feature selection. While for the other two modalities in chapters 3 and 4, there are relatively enough samples to use transform-learning-based models, Although developing deep-learning-based models in these modalities still requires more data, In addition, a number of labels are not covered in this research due to a lack of data.



## 5.3 Future Work

While each chapter focused on a specific aspect or modality, they all contributed to the development of a larger framework for analyzing parent-child interactions. In chapter 4, an ensemble model is proposed to combine audio and text modalities. Future work could explore the integration of these methods with other modalities such as video and face to further enhance performance. This work also could be extended to cover more labels in PCITS scale [17, 109] when enough number of samples are available. By providing more samples for the labels in the future, this system will be able to cover a larger number of labels for analysis. Using few-shot learning techniques could be another technique to incorporate more labels into the framework. An additional recommendation for future research involves concentrating on ethical matters related to the privacy of parents and their children. Implementing a federated learning strategy, which shifts computations to mobile devices, could be one possible solution to navigate these ethical concerns.

# Bibliography

- [1] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.
- [2] Parent-child relationship (pcrp). <https://barnardcenter.nursing.uw.edu/parent-child-relationship-programs-pcrp/>.
- [3] M. Buzzelli, A. Alb, and G. Ciocca, “A vision-based system for monitoring elderly people at home,” *Applied Sciences*, vol. 10, no. 1, p. 374, 2020.
- [4] S. M. Alghowinem, T. Gedeon, R. Goecke, J. Cohn, and G. Parker, “Interpretation of depression detection models via feature selection methods,” *IEEE Transactions on Affective Computing*, 2020.
- [5] G. K. Sidiropoulos, G. A. Papakostas, C. Lytridis, C. Bazinas, V. G. Kaburlasos, E. Kourampa, and E. Karageorgiou, “Measuring engagement level in child-robot interaction using machine learning based data analysis,” *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, ICDABI 2020*, 10 2020.
- [6] B. Huber, R. F. Davis, A. Cotter, E. Junkin, M. Yard, S. Shieber, E. Brestan-Knight, and K. Z. Gajos, “SpecialTime: Automatically detecting dialogue acts from speech to support parent-child interaction therapy,” *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, pp. 139–148, 5 2019. [Online]. Available: <https://dl.acm.org/doi/10.1145/3329189.3329203>
- [7] J. D. Domingo, J. Gomez-Garcia-Bermejo, and E. Zalama, “Improving human activity recognition integrating lstm with different data sources: Features, object detection and skeleton tracking,” *IEEE Access*, vol. 10, pp. 68 213–68 230, 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9807301>
- [8] S. Chen, J. Liu, H. Wang, and J. C. Augusto, “A hierarchical human activity recognition framework based on automated reasoning,” in *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2013, pp. 3495–3499.

- [9] S. Chen, K. Clawson, M. Jing, J. Liu, H. Wang, and B. Scotney, “Uncertainty reasoning based formal framework for big video data understanding,” *Proceedings - 2014 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2014*, vol. 2, pp. 487–494, 10 2014.
- [10] M. Ziaefard and R. Bergevin, “Semantic human activity recognition: A literature review,” *Pattern Recognition*, vol. 48, pp. 2329–2345, 8 2015, cite:212;pose-based methods;poselet-based methods: 62;SIFT, HOG, BoW;human pose estimation: 39;hierarchical: 102, 104, 105, 106, 107;semantic approaches;attribute based transfer learning: 35, 38;13.
- [11] R. Thomas, B. Abell, H. J. Webb, E. Avdagic, and M. J. Zimmer-Gembeck, “Parent-child interaction therapy: A meta-analysis,” *Pediatrics*, vol. 140, no. 3, 2017.
- [12] C. C. Lieneman, L. A. Brabson, A. Highlander, N. M. Wallace, and C. B. McNeil, “Parent-child interaction therapy: Current perspectives,” *Psychology research and behavior management*, pp. 239–256, 2017.
- [13] E. Batbaatar, M. Li, and K. H. Ryu, “Semantic-Emotion Neural Network for Emotion Recognition from Text,” *IEEE Access*, vol. 7, pp. 111 866–111 878, 2019.
- [14] M. B. Er, “A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features,” *IEEE Access*, 2020.
- [15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (almost) from Scratch,” *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 3 2011. [Online]. Available: <https://arxiv.org/abs/1103.0398v1>
- [16] V. Vajre, M. Naylor, U. Kamath, and A. Shehu, “PsychBERT: A Mental Health Language Model for Social Media Mental Health Behavioral Analysis,” *Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021*, pp. 1077–1082, 2021.
- [17] M. Oxford and D. Findlay, “Ncast caregiver/parent-child interaction teaching manual,” *Seattle, WA: NCAST Programs, University of Washington, School of Nursing*, 2013.

- [18] S. Alghowinem, H. Chen, C. Breazeal, and H. W. Park, "Body gesture and head movement analyses in dyadic parent-child interaction as indicators of relationship," *Proceedings - 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021*, 2021.
- [19] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *Acm Computing Surveys (Csur)*, vol. 43, no. 3, pp. 1–43, 2011.
- [20] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, pp. 30 509–30 555, 11 2020. [Online]. Available: <https://link.springer.com/article/10.1007/s11042-020-09004-3>
- [21] M. Khare and M. Jeon, "Multi-resolution approach to human activity recognition in video sequence based on combination of complex wavelet transform, local binary pattern and zernike moment," *Multimedia Tools and Applications*, pp. 1–30, 2 2022. [Online]. Available: <https://link.springer.com/article/10.1007/s11042-021-11828-6>
- [22] H. B. Zhang, Y. X. Zhang, B. Zhong, Q. Lei, L. Yang, J. X. Du, and D. S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors 2019, Vol. 19, Page 1005*, vol. 19, p. 1005, 2 2019. [Online]. Available: <https://www.mdpi.com/1424-8220/19/5/1005/htmlhttps://www.mdpi.com/1424-8220/19/5/1005>
- [23] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3d action recognition using learning on the grassmann manifold," *Pattern Recognition*, vol. 48, pp. 556–567, 2 2015.
- [24] J. Ye, G. Stevenson, and S. Dobson, "USMART," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 4, no. 4, 11 2014. [Online]. Available: <https://dl.acm.org/doi/10.1145/2662870>
- [25] S. Whitehouse, K. Yordanova, A. Paiement, and M. Mirmehdi, "Recognition of unscripted kitchen activities and eating behaviour for health monitoring," *IET Conference Publications*, vol. 2016, 2016.
- [26] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," 2015, pp. 4597–4605, cites: 582.
- [27] X. Ding, Q. Gan, and S. Bahrami, "A systematic survey of data mining and big data in human behavior analysis: Current datasets and models," *Transac-*

- tions on Emerging Telecommunications Technologies*, p. e4574, 2022. [Online]. Available: <https://onlinelibrary.wiley.com/doi/full/10.1002/ett.4574><https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.4574><https://onlinelibrary.wiley.com/doi/10.1002/ett.4574>
- [28] P. C. Roy, S. Giroux, B. Bouchard, A. Bouzouane, C. Phua, A. Tolstikov, and J. Biswas, “A possibilistic approach for activity recognition in smart homes for cognitive assistance to alzheimer’s patients,” in *Activity Recognition in Pervasive Intelligent Environments*. Springer, 2011, pp. 33–58.
- [29] K. Ramirez-Amaro, E. S. Kim, J. Kim, B. T. Zhang, M. Beetz, and G. Cheng, “Enhancing human action recognition through spatio-temporal feature learning and semantic rules,” *IEEE-RAS International Conference on Humanoid Robots*, vol. 2015-February, pp. 456–461, 2 2015.
- [30] B. Yao, H. Hagra, M. J. Alhaddad, and D. Alghazzawi, “A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments,” *Soft Computing*, vol. 19, pp. 499–506, 2 2015, cites: 64. [Online]. Available: <https://link.springer.com/article/10.1007/s00500-014-1270-4>
- [31] F. Wörgötter, A. Agostini, N. Krüger, N. Shylo, and B. Porr, “Cognitive agents—a procedural perspective relying on the predictability of object-action-complexes (oacs),” *Robotics and Autonomous Systems*, vol. 57, no. 4, pp. 420–432, 2009.
- [32] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes,” in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1778–1785.
- [33] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 951–958.
- [34] R. A. Güler, N. Neverova, and I. Kokkinos, “Densepose: Dense human pose estimation in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7297–7306.
- [35] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, “Rmpe: Regional multi-person pose estimation,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

- [36] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2d pose estimation using part affinity fields,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.
- [37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [38] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [39] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [40] D. Gong, G. Medioni, and X. Zhao, “Structured time series analysis for human action segmentation and recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1414–1427, 2013.
- [41] K. Li and Y. Fu, “Prediction of human activity by discovering temporal sequence patterns,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 8, pp. 1644–1657, 2014.
- [42] A. Grushin, D. D. Monner, J. A. Reggia, and A. Mishra, “Robust human action recognition via long short-term memory,” in *The 2013 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2013, pp. 1–8.
- [43] G. Lefebvre, S. Berlemont, F. Mamalet, and C. Garcia, “Blstm-rnn based 3d gesture classification,” in *International conference on artificial neural networks*. Springer, 2013, pp. 381–388.
- [44] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in neural information processing systems*, vol. 27, 2014.
- [45] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, “View adaptive neural networks for high performance skeleton-based human action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1963–1978, 2019.

- [46] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, “Deep multimodal feature analysis for action recognition in rgb+ d videos,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1045–1058, 2017.
- [47] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [48] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, “Language supervised training for skeleton-based action recognition,” 8 2022, <https://github.com/martinxm/lst>. [Online]. Available: <https://arxiv.org/abs/2208.05318v1>
- [49] W. Guilluy, L. Oudre, and A. Beghdadi, “Video stabilization: Overview, challenges and perspectives,” *Signal Processing: Image Communication*, vol. 90, p. 116015, 2021.
- [50] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [51] S. I. Serengil and A. Ozpinar, “Hyperextended lightface: A facial attribute analysis framework,” in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 2021, pp. 1–4. [Online]. Available: <https://doi.org/10.1109/ICEET53442.2021.9659697>
- [52] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [53] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “Yolov4: Optimal speed and accuracy of object detection,” *arXiv preprint arXiv:2004.10934*, 2020.
- [54] F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. L. Le, “Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection,” *Sensors*, vol. 21, no. 16, p. 5314, 2021.
- [55] P. Elias, J. Sedmidubsky, and P. Zezula, “Understanding the limits of 2d skeletons for action recognition,” *Multimedia Systems*, vol. 27, pp. 547–561, 6 2021. [Online]. Available: <https://link.springer.com/article/10.1007/s00530-021-00754-0>

- [56] K. Hara, H. Kataoka, and Y. Satoh, “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.
- [57] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.
- [58] E. V. Brestan and S. M. Eyberg, “Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids,” *Journal of clinical child psychology*, vol. 27, no. 2, pp. 180–189, 1998.
- [59] S. M. Eyberg, *Dyadic parent-child interaction coding system (DPICS): Comprehensive manual for research and training*. PCIT International, Incorporated, 2013.
- [60] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou, “Sentiment Embeddings with Applications to Sentiment Analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 496–509, 2 2016.
- [61] L. A. Cariola, S. Hinduja, M. Bilalpur, L. B. Sheeber, N. Allen, L.-P. Morency, and J. F. Cohn, “Language Use in Mother-Adolescent Dyadic Interaction: Preliminary Results,” *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, 10 2022. [Online]. Available: <https://ieeexplore.ieee.org/document/9953886/>
- [62] M. Kawamoto and A. Sashima, “Method for Analyzing Interactions in Pedagogical Environments using Environmental Sound Analysis,” *2021 2nd International Conference on Intelligent Data Science Technologies and Applications, IDSTA 2021*, pp. 54–59, 2021.
- [63] J. Xiao, Y. Huang, G. Zhang, and W. Liu, “A Deep Learning Method on Audio and Text Sequences for Automatic Depression Detection,” *Proceedings - 2021 3rd International Conference on Applied Machine Learning, ICAML 2021*, pp. 388–392, 2021.
- [64] T. Kwon, M. Jeong, E. S. Ko, and Y. Lee, “Captivate! Contextual Language Guidance for Parent-Child Interaction,” *Conference on Human Factors in Computing Systems - Proceedings*,



vol. 17, 4 2022. [Online]. Available: <https://dl.acm.org/doi/10.1145/3491102.3501865>

- [65] E. Hossain, M. L. Cahoon, Y. Liu, C. Kurumada, and Z. Bai, "Context-responsive ASL Recommendation for Parent-Child Interaction," *ASSETS 2022 - Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, vol. 5, no. 22, p. 2022, 10 2022. [Online]. Available: <https://dl.acm.org/doi/10.1145/3517428.3550366>
- [66] I. Hwang, C. Yoo, C. Hwang, D. Yim, Y. Lee, C. Min, J. Kim, and J. Song, "Talkbetter: family-driven mobile intervention care for children with language delay," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 2014, pp. 1283–1296.
- [67] S. Song, S. Kim, J. Kim, W. Park, and D. Yim, "Talklime: mobile system intervention to improve parent-child interaction for children with language delay," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 304–315.
- [68] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015," 9 2015. [Online]. Available: <https://repositories.lib.utexas.edu/handle/2152/31333>
- [69] J. T. Hancock, C. Landrigan, and C. Silver, "Expressing emotion in text-based communication," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 929–932, 2007. [Online]. Available: <https://dl.acm.org/doi/10.1145/1240624.1240764>
- [70] B. Shickel, S. Siegel, M. Heesacker, S. Benton, and P. Rashidi, "Automatic Detection and Classification of Cognitive Distortions in Mental Health Text," *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*, pp. 275–280, 10 2020.
- [71] J. Gilkerson, J. A. Richards, S. F. Warren, J. K. Montgomery, C. R. Greenwood, D. K. Oller, J. H. Hansen, and T. D. Paul, "Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis," *American Journal of Speech-Language Pathology*, vol. 26, no. 2, pp. 248–265, 5 2017. [Online]. Available: [https://pubs.asha.org/doi/full/10.1044/2016\\_AJSLP-15-0169](https://pubs.asha.org/doi/full/10.1044/2016_AJSLP-15-0169)
- [72] I. Hwang, C. Yoo, C. Hwang, D. Yim, Y. Lee, C. Min, J. Kim, and J. Song, "TalkBetter: Family-driven mobile intervention care for children with language delay," *Proceedings of the ACM*

- Conference on Computer Supported Cooperative Work, CSCW*, pp. 1283–1296, 2014. [Online]. Available: <https://dl.acm.org/doi/10.1145/2531602.2531668>
- [73] W. Yin, K. Kann, M. Yu, and H. Schütze, “Comparative study of CNN and RNN for natural language processing,” *arXiv preprint arXiv:1702.01923*, 2017.
- [74] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 10 2018. [Online]. Available: <https://arxiv.org/abs/1810.04805v2>
- [75] A. Younus and M. A. Qureshi, “Combining BERT with Contextual Linguistic Features for Identification of Propaganda Spans in News Articles,” *Proceedings - 2020 IEEE International Conference on Big Data, Big Data 2020*, pp. 5864–5866, 12 2020.
- [76] R. Sennrich, B. Haddow, and A. Birch, “Improving Neural Machine Translation Models with Monolingual Data,” *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, vol. 1, pp. 86–96, 11 2015. [Online]. Available: <https://arxiv.org/abs/1511.06709v4>
- [77] S. Yoosuf, Y. . David, and . Yang, “Fine-Grained Propaganda Detection with Fine-Tuned BERT,” pp. 87–91, 11 2019. [Online]. Available: <https://aclanthology.org/D19-5011>
- [78] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” *ACM International Conference Proceeding Series*, vol. 382, 2009. [Online]. Available: <https://dl.acm.org/doi/10.1145/1553374.1553380>
- [79] S. Dessai and S. S. Usgaonkar, “Depression Detection on Social Media Using Text Mining,” *2022 3rd International Conference for Emerging Technology, INCET 2022*, 2022.
- [80] N. S. Alghamdi, “Monitoring mental health using smart devices with text analytical tool,” *2019 6th International Conference on Control, Decision and Information Technologies, CoDIT 2019*, pp. 2046–2051, 4 2019.

- [81] Q. T. Nguyen, T. L. Nguyen, N. H. Luong, and Q. H. Ngo, “Fine-Tuning BERT for Sentiment Analysis of Vietnamese Reviews,” *Proceedings - 2020 7th NAFOSTED Conference on Information and Computer Science, NICS 2020*, pp. 302–307, 11 2020.
- [82] D. Goularas and S. Kamis, “Evaluation of deep learning techniques in sentiment analysis from twitter data,” in *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*. IEEE, 2019, pp. 12–17.
- [83] Y. Chen, “Convolutional Neural Network for Sentence Classification,” 8 2015. [Online]. Available: <https://uwspace.uwaterloo.ca/handle/10012/9592>
- [84] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for Word Representation,” pp. 1532–1543. [Online]. Available: <http://nlp>.
- [85] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” *7th International Conference on Learning Representations, ICLR 2019*, 11 2017. [Online]. Available: <https://arxiv.org/abs/1711.05101v3>
- [86] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [87] “BERT,” [https://huggingface.co/docs/transformers/model\\_doc/bert](https://huggingface.co/docs/transformers/model_doc/bert).
- [88] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, “Captum: A unified and generic model interpretability library for PyTorch,” 9 2020. [Online]. Available: <https://arxiv.org/abs/2009.07896v1>
- [89] B. Liu, “Sentiment analysis and opinion mining,” *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [90] C. Strapparava and R. Mihalcea, “Learning to identify emotions in text,” in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560.

- [91] R. Plutchik, “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice,” *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.
- [92] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 10 2014.
- [93] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1746–1751, 2014. [Online]. Available: <https://aclanthology.org/D14-1181>
- [94] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [95] S. Xu, Z. Yang, D. Chakraborty, Y. Tahir, T. Maszczyk, V. Y. H. Chua, J. Dauwels, D. Thalmann, N. M. Thalmann, B. L. Tan, and J. L. C. Keong, “Automatic Verbal Analysis of Interviews with Schizophrenic Patients,” *International Conference on Digital Signal Processing, DSP*, vol. 2018- November, 1 2019.
- [96] N. Wang, M. Chen, and K. P. Subbalakshmi, “Explainable CNN-attention Networks (C-Attention Network) for Automated Detection of Alzheimer’s Disease,” 6 2020. [Online]. Available: <https://arxiv.org/abs/2006.14135v2>
- [97] D. Gordeev, “Detecting state of aggression in sentences using cnn,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9811 LNCS, pp. 240–245, 2016. [Online]. Available: [https://link.springer.com/chapter/10.1007/978-3-319-43958-7\\_28](https://link.springer.com/chapter/10.1007/978-3-319-43958-7_28)
- [98] N. T. Vu, H. Adel, P. Gupta, and H. Schütze, “Combining recurrent and convolutional neural networks for relation classification,” *arXiv preprint arXiv:1605.07333*, 2016.
- [99] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, “Deep features-based speech emotion recognition for smart affective services,”

- Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5571–5589, 3 2019. [Online]. Available: <https://link.springer.com/article/10.1007/s11042-017-5292-7>
- [100] T. Anvarjon, Mustaqeem, and S. Kwon, “Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features,” *Sensors 2020, Vol. 20, Page 5212*, vol. 20, no. 18, p. 5212, 9 2020. [Online]. Available: <https://www.mdpi.com/1424-8220/20/18/5212/html>  
<https://www.mdpi.com/1424-8220/20/18/5212>
- [101] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 86–96. [Online]. Available: <https://aclanthology.org/P16-1009>
- [102] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozire, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [103] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [104] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [105] G. King and L. Zeng, “Logistic regression in rare events data,” *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001.
- [106] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [107] B. Nikbakhtbideh, L. Duffett-Leger, and M. Moshirpour, “Behavior analysis of parent-child interactions from text,” *Proceedings - 22nd International Conference on Machine Learning and Applications, ICMLA*, 2023, to be published.

- [108] M. Kanakaraj and R. M. R. Guddeti, "Nlp based sentiment analysis on twitter data using ensemble classifiers," in *2015 3Rd international conference on signal processing, communication and networking (ICSCN)*. IEEE, 2015, pp. 1–5.
- [109] G. Sumner, "Ncast caregiver/parent-child interaction teaching manual. seattle," *NCAST Publications*, 1994.

# Appendix A

## PCITS Labels

Label	Description	Actors	Modalities
Back arching	Hyperextension of the back.	Child/Parent	Image
Crawling away	Moving away from the caregiver on hands and knees.	Child/Parent	Image
Maximal lateral gaze aversion	Maximum turning of head laterally accompanied with gaze aversion; may have slight back arching.	Child/Parent	Image
Overhand beating movements	Of arms, elbows flexed tightly, upper arms raised, hands fisted at shoulder level, then the arm is brought straight down. Throwing objects would be overhand beating movements.	Child/Parent	Image
Pale/red skin	Skin changes colour to either pale or red.	Child	Image
Pulling away	Removing the torso and/or head away from the caregiver or object; withdrawing and increasing distance from the caregiver or object.	Child/Parent	Image
Pushing away	Making manual contact with the caregiver or object and extending the arm.	Child/Parent	Image
Tray pounding	Hitting a surface such as a high chair tray or tabletop with the palm of the hand.	Child/Parent	Image
Nonverbal Soothing	pat, rock, kiss, touch	Child/Parent	Image
Head Nod	Head up and down	Parent	Image
Imperative Instruction	A given instruction or command (e.g., "Catch it.")	Parent	Text/Audio
Continued on next page			

**Table A.1 – continued from previous page**

<b>Label</b>	<b>Description</b>	<b>Actors</b>	<b>Modalities</b>
Explanatory In- struction	A given instruction to explain why something needs to be done a certain way (e.g., “I want to see if you can turn the page.”).	Parent	Text/Audio
Broad Praise	Indicates general praise given (e.g., “I’m proud of you”).	Parent	Text/Audio
Task-Related Praise	Indicates praise given by the adult to the child specifically related to the task (e.g., “Good job. Did it”).	Parent	Text/Audio
Cheerleading	Indicates a statement of encouragement or motivation (e.g., “You can do it too.”).	Parent	Text/Audio
Negative Comment	Indicates a negative comment or feedback (e.g., “Open your hand and Oh nope.”).	Parent	Text/Audio



## Appendix B

### Video Tracking Feature Extraction Algorithm

---

**Input :**  $frames = [F_i], i = 1, 2, \dots, N$ ; each  $F_i$  contains  $persons = [P_j]$  and  $keypoints = [KP_k]$   
**Output:**  $track = [KPC_i, KPP_i]$ , where each  $KPC_i$  contains the normalized state of keypoints of the child relative to the parent and each  $KPP_i$  contains the relative position of parent's keypoints to the self head.

```
1 for frame in frames:
2 // adjust coordinates of persons
3   frame.persons = adjust_coordinates(frame.persons)
4 // count number of joints for each person
5   count_joints = count_joints(frame.persons, frame.keypoints)
6 // remove persons without any joint
7   frame.persons = filter(if_has_any_joint, frame.persons, count_joints)
8 // select persons with more joints if there are more than 2 persons in the frame
9   if len(frame.persons) > 2:
10    frame.persons = sublist(sorted(number_of_joints, frame.persons), 2)
11 // remove frames without any detected persons
12   elif len(frame.persons) == 0:
13    frames.remove(f)
14 // the same filters for keypoints
15 // cluster all detected persons into two groups
16 clustered_persons = cluster([p.left, p.right, p.top, p.bottom] for p in frame.persons) for frame in frames)
17 // calculate average area of each cluster, assign smallest as the child
18 avg = [average(calculate_area, clustered_persons[i]) for i in [0, 1]]
19 label_child_1 = min_index(avg)
20 // cluster all detected keypoints into two groups
21 clustered_keypoints = cluster([kp.points for kp in frame.keypoints] for frame in frames)
22 // find the keypoints cluster index that covers more persons with type = label_child_1
23 label_child_2 = max(coverage(clustered_keypoints, clustered_persons[label_child_1]))
24 // match keypoints and persons by their cluster index
25 track = [(p_child, kp_child, p_parent, kp_parent) for (p_child, kp_child, p_parent, kp_parent) in
26   zip(clustered_persons[label_child_1], clustered_keypoints[label_child_2],
27     clustered_persons[not(label_child_1)], clustered_keypoints[not(label_child_2)])]
28 for (p_child, kp_child, p_parent, kp_parent) in track:
29 // for keypoints without any person, estimate it based on next/previous frames
30   if p_child is None:
31     p_child = average(kp_child.frame.next.p_child, kp_child.frame.prev.p_child)
32 // remove joints outside person bbox
33 kp_child = filter(is_inside(p_child), kp_child)
34 // do the same for p_parent, kp_parent
35 // transform by parent's head position
36 kp_child = kp_child - kp_parent.head
37 // normalize by image width/height
38 kp_child = kp_child / frame_dimension
39 track_child = [kp_child for (_, kp_child, _) in track]
40 // similarly do for parents
41 return [track_child, track_parent]
```

# Appendix C

## Labelling System

Trained research assistants are responsible for preparing and labelling the videos using a user interface developed in React.js and Node.js and deployed on Amazon Web Services (AWS) using Amplify <sup>1</sup>, DynamoDB <sup>2</sup>, and S3 <sup>3</sup>. Figure C.1 shows this interface. The requirement gathering of the labelling system was done through several meetings with the domain experts in nursing department of University of Calgary according to the PCITS scale [17, 109]. The labelling tool was initially created in this research for nurse researchers and their team members to label mother-infant interaction videos. This system helped gather the required samples for AI analysis. Each sample contains the name of the video file, the label, and the start and finish times of the event. Another coder reviews these samples and confirms the video as “labelled”. The videos are stored in AWS S3 and the labels are preserved in DynamoDB. The development and deployment environments were handled by AWS Amplify. In this research, the required samples were fetched from AWS S3 and DynamoDB. A number of lambda functions were developed for data pipelining and preparation, including a service that converts video files to audio and frames images by utilizing the FFmpeg library in Python. The other service transcribes the audio files using the AWS Transcriber service. The other service was utilized for data preparation for the video modality and also relevant lambda functions for audio and text modalities. The line of codes for the implementation of Python codes is 38000, and for the labelling system, it is 5000. The code of the AI modules is maintained in <https://github.com/aranite-open/vidkids-ai> and for the labelling system in <https://github.com/aranite-open/vidkids-ai-labeling>.

---

<sup>1</sup><https://aws.amazon.com/amplify/>

<sup>2</sup><https://aws.amazon.com/dynamodb/>

<sup>3</sup><https://aws.amazon.com/s3/>

Video Labeling [Videos](#) [Upload](#) [Sign Out](#)

**VIDEOS**

Search by Name

Labeled  Unlabeled

4

VK519-Mar22-2022-Visit#2.mp4

VK117-Mar18-2021-V2.mp4

VK042-Jan.29-2019-V4.MP4


VK176 Visit#2 July 6, 2022.mp4

VK124 Dec312020 V2 zoom.mp4

VK127-Nov24-2020-Visit 2.MP4

**Default Quality** **Medium Quality** **Low Quality**

Filename: ce96ef76-fc5c-41dd-b77e-6a7b6396cb6e-VK042-Jan.29-2019-V4.MP4



0:23 / 2:58

**Disengagements**

1.	Pulling away	<input type="checkbox"/>
From:	00:22.994	<input type="checkbox"/>
To:	00:23.014	<input type="checkbox"/>
2.	Tray pounding	<input type="checkbox"/>
From:	00:46.983	<input type="checkbox"/>
To:	00:47.495	<input type="checkbox"/>
3.	Whining	<input type="checkbox"/>
From:	01:37.899	<input type="checkbox"/>
To:	01:39.422	<input type="checkbox"/>
4.	Overhand beating movements	<input type="checkbox"/>
From:	01:46.851	<input type="checkbox"/>
To:	01:47.735	<input type="checkbox"/>
5.	Crawling away	<input type="checkbox"/>
From:	02:33.853	<input type="checkbox"/>
To:	02:35.815	<input type="checkbox"/>
6.	Overhand beating movements	<input type="checkbox"/>
From:	02:55.221	<input type="checkbox"/>
To:	02:56.865	<input type="checkbox"/>

Figure C.1: Labelling system

# Appendix D

## Copyright Permissions

In this thesis, I used three datasets. Two datasets related to Chapter 2 and Chapter 4 are collected from the data in the labelling system C, and one dataset related to the Chapter 2 is downloaded from the publicly available dataset at <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/C5Z3SC> with licence CC0 1.0 which implies “No Copyright”.

Below is the confidentiality agreement form that was signed before starting to use parent-child interaction data in this research.



## EMPLOYEE CONFIDENTIALITY AGREEMENT

The University of Calgary places a high level of responsibility and trust in its employees especially those who handle human resources data. As part of your duties, you have access to confidential records and information regarding the University, its employees, consultants and agents. With respect to these records and information, and all other confidential and proprietary University of Calgary information and records, I agree to the following:

1. I acknowledge the confidentiality of all employee information and records and other confidential and proprietary University information and records. I agree that this information will not be revealed to or distributed to or discussed with anyone other than the appropriate, designated supervisor or other University officials.
2. I will not attempt to alter, change, modify, add, or delete employee record information or University documents unless doing so is part of my assigned job duties.
3. I will access only the information specified and authorized by my supervisor. Access to information should be through normal office procedures for obtaining specific access to the information in written documents, computer files, or other University information.
4. I understand that failure to abide fully by the above agreement is grounds for immediate discipline, up to and including termination of employment.

The University of Calgary reserves the right to perform access audits of the system periodically and without prior notification.

Employee/Name (Printed)

Behnam Nikbakhsh

Employee/Signature and Date

[Redacted Signature]

2021, 07, 23

Supervisor

Signature and Date

\_\_\_\_\_



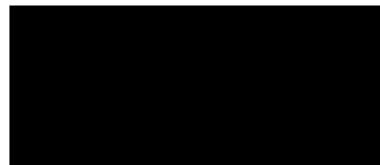
Confidentiality Agreement for Research Assistants / Transcribers/Translators

Name of Researcher: Dr. Nicole Letournau

Title of Project: CHILD Studies

Before we can hire you to transcribe research interviews, we must obtain your explicit consent not to reveal any of the contents of the tapes, nor to reveal the identities of the participants (i.e. the students and supervisors interviewed and their place of employment). If you agree to these conditions, please sign below.

Behnom Nikbakhtchi  
Print Name



Signature  
2021, 07, 23