Exploring Emotion Recognition in Conversational Data: A Focus on Explainability

Knut Stautland Ivarsøy

Informatikk: programmering og systemarkitektur,
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

Spring 2023

Abstract

The development of an avatar-based interview training system to enhance police officers' ability to retrieve information is a difficult task. It involves integrating various components, such as an emotional component. Using deep-learning neural networks (DNN) machines can understand and express emotions. However, a problem with DNNs is the lack of transparency and interpretability. This work proposes to use explainable AI on a DNN to explain sentiment profiles generated by language models. Moreover, we explore the explanations to unveil patterns, sentences, or sections typical for abusive transcripts. We applied language models on abusive and non-abusive transcripts to generate sentiment profiles of emotion confidence scores. These sentiment profiles are then inserted into a convolutional neural network (CNN) for classification. With the classification of the CNN, we use Local Interpretable Model-Agnostic Explanations (LIME) to extract the key features of the predictions. We present them in a heat map for visualization and analysis. Analysis of the heat maps generated from abusive sentiment profiles displayed clusters of important features for the emotions of fear, anger, and disgust. In the heat maps generated from non-abusive transcripts, no notable patterns or sections were identified. One possible explanation is that the non-abusive transcripts are a combination of various transcript types grouped under a single class, the non-abusive class. In conclusion, we see that by using two language models to generate sentiment profiles the CNN's certainty in important emotions increases.

Acknowledgments

I would like to thank my supervisors Pål Halvorsen and Michael Riegler. Additionally, I would like to thank my informal supervisor Syed, Zohaib Hassan. Without their expertise and guidance, I would never have been able to finish this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

When police officers are interviewing alleged victims of abuse, it is essential to show empathy, especially if the victim is a minor [1], [2]. In these situations, the child is often the only source of information for investigators [1]. It has been shown that children can be competent informants in conversations [1]. Acquiring the skills to conduct high-quality interviews could take years to master.

AI-Based avatar project is an ongoing collaboration between OsloMet, SimulaMet, Child protection services, and several international partners [3], [4], [5]. The objective is to generate an AI-Based child avatar. The reason for building an AI-Based avatar is to offer police officers or other law personnel the chance to increase their child interviewing skills dynamically [3]. Studies have shown that the standard classroom teaching model is not the most effective training model [6]. Furthermore, the classroom teaching model also requires a schedule and a teacher. The AI-Based avatar would provide law enforcement with individual feedback and the ability to practice on a range of contexts. In the article by Benson and Powell [6], one of the participants said, "I really liked the online course because it was always there and you can do it whenever you're ready."

Law enforcement could practice on the avatar whenever their schedule is open without relying on others. Being dependent on a teacher's schedule could cause the participants to deprioritize strengthening their skills in conducting informative interviews.

This thesis will extract sentiment scores from transcripts using existing language models and use these scores as input into a deep neural network (DNN). Finally, the DNNs' decisions will be analyzed utilizing an explainable AI python library for extracting important features. We will examine what affects the DNN's decision and if important features give information about important sentences, words, or sections in an abusive transcript. This will be done by generating a heatmap of the important features. Inspecting the heatmap could reveal how a conversation with a child who is an alleged victim of abuse is different than with other children.

Determining if a transcript is abusive or not abusive requires the full context of the conversation. A transcript could express abusive behavior when only inspecting a specific section, but when grasping the full context, it could be revealed to be a transcript about something else. Therefore, this experiment will also try to understand if it is possible to detect the type of transcript using DNNs.

1.2 Problem statement

DNNs can be labeled "black boxes." They receive an input, compute the values of that input and then return an output. They excel at complex tasks such as image classification and text generation. However, can the predictions made by DNNs be trusted? The prediction of a DNN often returns the output without explaining why. There have been several attempts to solve this problem [7], [8]. Trusting something often relies on understanding why and how. In this thesis, we explore using explainable AI to analyze the prediction of a DNN with sentiment profiles generated by language models as input. This will be answered by the following research

questions

1. Can explainable AI be used to uncover the decisions in deep learning models?

2. Are there certain patterns of dialogue within abusive and non-abusive interviews?

3. Can emotion analysis be used to classify the type of transcript?

We aim to determine whether explainable AI can identify the important features in the DNN's predictions. Additionally, we intend to explore whether these features can help differentiate between an abusive and non-abusive transcript.

1.3 Scope and limitations

The scope of this thesis is within the AI-based avatar project. Therefore, the systems and transcripts used will be based on the context of children. The aim is to uncover the decisions made by a DNN and use them for better expression and understanding of emotions. This is accomplished through the use of sentiment analysis by language models, which are benchmarked by testing their ability to classify children's emotions. The language models then create sentiment profiles of confidence scores for two separate datasets - one containing abusive interviews with children and the other non-abusive interviews with children. These sentiment profiles are then inserted into a DNN in order to analyze the prediction. To extract the important features behind the prediction we used an explainable AI python library called LIME [8].

We noticed two especially crucial limitations in reference to the scope of the thesis, the abusive datasets and the available language models. As abusive transcript with children is not something made public, it was difficult to gather enough samples. The abusive dataset gathered consisted only of 20 transcripts. Additionally, these transcripts contained fewer utterances

than the non-abusive transcripts causing the sentiment profiles to require manipulation. A last limitation regarding the datasets was that the non-abusive transcripts contained a lot of information that makes sense for other research fields rather than language models, including typographical symbols and written gestures. This made the task of making the non-abusive dataset ready for use quite demanding and difficult.

The language model limitations revolved around available language models for sentiment analysis on children. Most language models created are trained on datasets with adult conversations. However, children and adults do not always articulate similarly. A solution could be to fine-tune a model on dialogues including children. However, as we did not have enough abusive transcript, this would be difficult.

1.4 Research method

In the paper by Denning et al. [9], Association for Computing Machinery (ACM) Task Force presented a set of steps to follow when conducting research. It consists of three paradigms, theory, abstraction, and design [10]. This thesis research takes advantage of all these paradigms. Gathering information about the available language models and datasets were based on the theory paradigm. Benchmarking the language models combined with creating a visualization for extracted features was based on the abstraction and design paradigm.

Additionally, we utilized both qualitative and quantitative research approaches [11]. A quantitative approach was used when gathering the results to detect patterns within transcripts. Then we used a qualitative approach to explain why we saw these and the sentences representing these features.

1.5 Ethical considerations

In a broad setting, language models could create their own social norms. For example, it could automatically label a unisex name to a specific gender. It could assume a family must consist of a father, mother, and child/children. And as will be discussed in this thesis, it could predict wrong emotions and misbehave. Most can attest that understanding another person's feelings or mental state can be complex. Therefore, why should we trust AI to understand emotional intelligence better? In the paper by Stark and Hoey [12], they argued that more science, laws, and technologies are needed when designing AI systems concerned with artificial emotional intelligence.

Another ethical consideration is that language models require data for training, which may include sensitive information. As demonstrated by the release of chat-GPT to the public, language models are vulnerable to manipulation for retrieving information. Therefore, there is a need to ensure privacy and that appropriate safety measures are in place. Both because it is by law, but also to protect abused children's information from bad actors.

A more specific ethical consideration for this thesis is deciding whether a transcript is abusive or not purely based on emotions. Mislabeling an abused transcript solely on emotion could cause damage if acted upon.

1.6 Main contributions

In this thesis, we explore the important features behind a DNNs decision in the context of using an AI-based child avatar. We did this by analyzing the predictions made by a convolutional neural network (CNN) on language models' predicted sentiment profiles. Each sentiment profile consists of confidence scores displaying the emotional state of a child throughout the transcript. We extracted the important features using an explainable AI python library. The important features extracted were then visualized in a

heat map for comparison and analysis.

We had three research questions we wanted to answer, see section 1.2. LIME [8] was utilized to extract the features from the prediction. In addition to extracting important features, LIME [8] assigned an importance value to each feature. This, together with patterns revealed certain emotions important behind the prediction. From analyzing the heat maps we noticed repeating patterns and groups of features across the heat maps. While this simply indicates that there is a high emotion value in a certain region, we saw that when using two language models sentiment profiles instead of one, these groups and values became more established. Additionally, inspecting the sentences representing the features often revealed the abusive content of the transcripts.

As to the question of if explainable AI can differentiate between abusive and non-abusive transcripts, we noticed that the non-abusive features did not reveal any particular patterns. The main impression gained from the non-abusive heat maps was the absence of abusive feature groups. However, we argue that this is expected as it is difficult to discover features that do not represent something. The prediction score for the CNN displayed an accuracy of 100%. Combining this with the fact that the feature extracted revealed information, we conclude that there is some trust gained in the prediction.

1.7 Thesis Outline

Chapter 2 - Background Chapter 2 introduces important theoretical information about technologies utilized in the thesis. It starts with a general overview and challenges of the AI-based avatar project. Then, introduces the concept of predicting and expressing emotions. We end the chapter by exploring models and datasets relevant to the thesis.

Chapter 3 – Design and Implementation Chapter 3 discusses the steps executed for explainability analysis. Starting with discussing the datasets

utilized for both benchmarking and the creation of sentiment profiles. Then, we introduce the benchmarking process, where the ability of the language models' are tested within the scope of the thesis. Using the capable models we explain how they were applied to creating sentiment profiles. Finally, we explain the generating of heat maps used for explainability analysis.

Chapter 4 - Results Chapter 4 introduces examples of heat maps and analyses the observations made. We start by presenting the metric used, heat maps, and relevant information about them. Then we investigate the heat maps generated from the language models' sentiment profiles one by one. The chapter ends by comparing two abusive heat maps from the same transcript.

Chapter 5 - Discussion This chapter explains the analysis of the heat maps in a broader setting. It also discusses valuable information and non-valuable information gained from the feature extraction. It ends with describing challenges and future experiments that can be performed for better results.

Chapter 6 - Conclusion Chapter 6 concludes the thesis by summarizing what was done. Future work is proposed to increase the understanding of the problem statement further.

Chapter 2

Background

2.1 Virtual Avatars for Investigative Interviews with Children

Child abuse is a major problem in the world that could hinder a child's development and impact the child's current situation [3]. The abused victims are often the only reliable source of information [13]. Therefore, it is important to give them an opportunity to speak. Research into the field of conducting informative investigative interviewing on children has been developed for several years [5]. This research has shown that using open-ended questions achieves better results [14]. In Baugerud et al. [15], Baugerud et al. argue that previous studies from multiple countries showed a lack of quality in investigative interviews. A proposed solution to this is using a virtual avatar for investigative interview training. A joint project between OsloMet, SimulaMet, Child protection services, and several international partners proposes a training program to improve investigative interviewing [14]. The proposed avatar consists of several different components combined; chatbots, visual content, text-to-speech, and speech-to-text [14].

This thesis will focus on the textual component of the avatar, more

specifically the emotional intelligence of the avatar. "The results of an investigative interview with an adult depend on how officers handle the emotions of the interviewee", in Hassan et al. [5]. Emotions play an integral part when conducting investigative interviewing. Using emotions can help interviewees to retrieve more information or enhance their memory [16]. In figure 2.1, the yellow boxes represent the textual components in the avatar. Both facial expressions and language need to express emotions. Several technologies show positive results in the field of understanding and expressing emotions. A few of these technologies will be discussed later in the thesis.

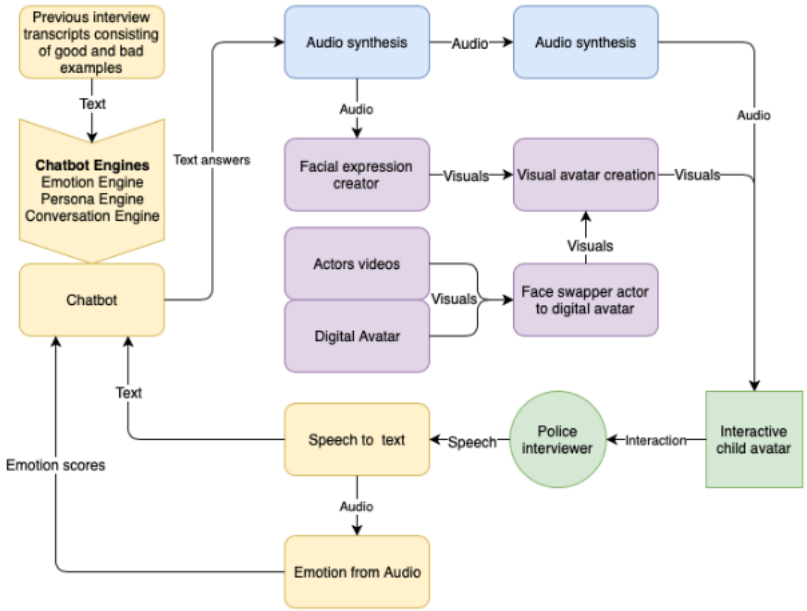

Figure 2.1: Talking Child-Avatar architecture [14]

2.2 Short introduction to chatbots

The English dictionary Lexico [1] defines a chatbot as "A computer program designed to simulate conversation with human users, especially over the Internet". Chatbots are used to communicate with people. They

[1] https://www.lexico.com/en/definition/chatbot

are popular in various applications like education, health, and customer service. By using a chatbot, companies can help multiple customers at a time and thereby save both money and time.

A problem with chatbots is mimicking human emotions, and acting like human beings. A study conducted by Luo et al. [17], showed that when a customer realized the other person was a chatbot, the chance of selling a product decreased by 79,7%. Humans do not view talking to a chatbot, as the same as talking to another human being. If the problem with speaking with chatbots is that they do not act like a human. A solution could be creating a more emotionally aware chatbot. Emotions expressed by a chatbot might affect the interaction between the human and the chatbot, for example in educational, supportive, or other interactions. There have been progressions in recent years in designing chatbots that can express human-like emotions [18].

2.2.1 Pattern matching approach

The two main approaches when developing a chatbot are either pattern-matching or using Artificial Intelligence Markup Language (AIML) [19]. The pattern-matching approach is based on rule patterns. Moreover, it matches the user's input to a set of rules. Then, it selects what to return as a response based on the pattern-matching algorithm implemented. A flaw with this approach is the chatbot patterns often are handmade. Therefore, scaling the chatbot would require a lot of work [20].

AIML is an open-source pattern-matching approach. AIML is built on data objects, which contain topics and categories. The topics contain categories, and each of these categories matches the input to an output. The output generated is based on the template in the data object [21].

2.2.2 Natural Language Processing

Natural Language Processing (NLP) helps the computer understand sentences written in natural languages [22]. This is useful for machine-

learning chatbots because they need a method to convert user input into readable data for the machine. NLP can be divided into two categories called natural language understanding (NLU) and natural language generation (NLG) [19]. NLU decides the meaning of an input, using syntactic and semantic analysis [23]. NLG determines the response to a certain input. The main difference between a pattern-matching method and NLP is that NLP focuses on the context of the conversation instead of only single inputs.

Nowadays, NLP and Machine Learning (ML) are used together [24]. To determine the response using a machine-learning model, the popular choice is Artificial Neural Networks (ANNS). Neural networks work by assigning scores to generate responses and then choosing the response with the highest score. The scores are achieved through training the neural network on a dataset. Choosing the correct dataset can prove to be difficult because the training set might be too big or small, it could lack normal everyday conversations, or it could have several grammatical errors [25]. In the next sections, we will explain datasets and neural networks in more detail.

2.3 Machine learning

Machine learning is, as its name tells, creating methods for machines to learn. There are four machine-learning approaches; Supervised, unsupervised, semi-supervised, and reinforcement - learning. In the following section, we examine concepts in machine learning.

2.3.1 Deep learning

Deep learning is a neural network method for solving tasks without human intervention. Utilizing deep learning, we can eliminate the pre-processing of data. Different layers in the DNNs represent various features in the data [26]. These layers work together to extract key features from the input

data. For example, when deciding whether a picture is a car or not. A deep learning algorithm can extract the important features of a car, such as wheels.

2.3.2 Neural Network

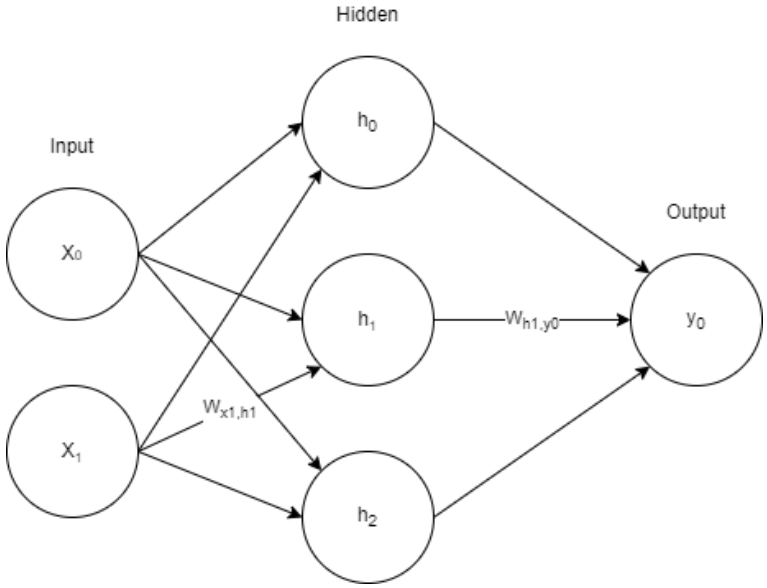

Figure 2.2: Neural network with a single hidden layer

Neural networks (NN) consist of "neurons" connected to act like a human brain. They are connected in layers, the example seen in figure 2.2 shows the architecture of a single-layer NN. Each neuron within these layers has weight and bias values for computation to manipulate the input. A neuron in a NN is called a perceptron. There are several NN architectures; this thesis will inspect recurrent networks and convolutional neural networks (CNN).

Learning within a neural network consists of updating the weights between the neurons. There are three main learning techniques; supervised, unsupervised, and a hybrid method. Supervised adds the correct output value to the input. In contrast, unsupervised does not add this value.

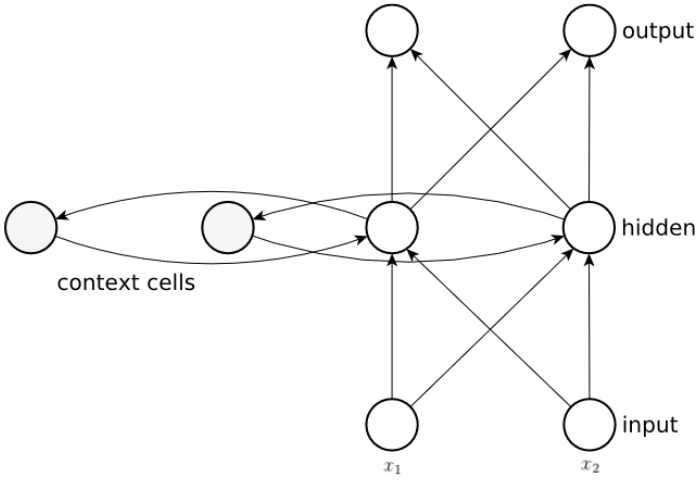

Figure 2.3: A recurrent neural network [27]

2.3.3 Recurrent Neural Network

Recurrent neural network (RNN) architecture has a circular connection between the neurons, as shown in figure 2.3. When training, the network propagates data backward [27]. This achieves memory within the network. The output of the hidden layer is saved in context cells, which each neuron possesses. There are two main learning approaches for RNNs. These two are backpropagation through time (BPTT) and real-time recurrent learning (RTRL). Calculation of the weights is the main difference between these two.

BPTT uses backpropagation to update the weights after calculating the training sequence [27]. This is done by unfolding the network into a feed-forward neural network that would exist at that time unit. For each time unit, the weights are updated. RTRL, however, updates the weight while receiving the input. A consequence of this is a high computational cost.

2.3.4 Convolutional Neural Network (CNN)

CNN is a neural network that excels in the field of computer vision [28]. Examples are image processing, pattern recognition, and voice recognition.

The difference between a CNN and other artificial neural networks is that a CNN is not concerned with where a pattern is located. It is concerned with recognizing the pattern. This is often done with layers, the first layer might detect a certain feature, while other layers detect other features.

Architecture

The architecture of CNN often varies in the number of layers and operations applied. A constant in all CNN's is the convolution layer. The convolution layer is often the first layer, which convolves the image. A simple CNN could have the architecture:

$$Convolution\ layer \rightarrow ReLU \rightarrow Max\ pool \rightarrow Fully\ connected\ layer \quad (2.1)$$

The output size from the convolution layer depends on the kernel size and stride. A kernel is a filter applied to the picture to decrease the size of the required weights and causes each neuron to analyze a specific image region. Stride is a component that determines the shift of the kernel across the image [29]. The following sections will explain these layers in more detail.

Convolution Layer

The convolution layer receives pixels as input. If a picture is 64x64 pixels, every pixel would be an input [28], [29]. The number of weight connections required will be huge if every neuron in the convolution layer is connected to every pixel. Therefore, a solution in CNNs is using filters to decrease these connections. This means that different layers could look at different features in an image. The filter is called a kernel. This kernel is applied to the entire picture, pixel by pixel. Stride and the kernel size select how much of the picture to inspect at each iteration. Stride is how far the kernel should move by each iteration. Stride = 1 would turn an input image size = 5x5 with kernel size 3x3 to a 3x3 output image. An example shown in figure 2.4, the kernel is applied to the top 3 pixels 3 times, then moved down one pixel. It would do this until all pixels are covered at least once.

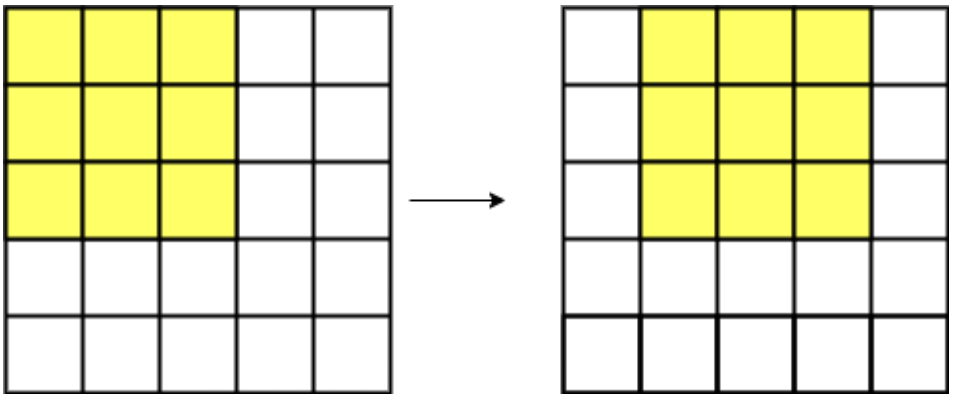

Figure 2.4: CNN with kernel size 3 and stride 1

Pooling layer

There are usually two alternatives in this layer, max pooling, and L2 pooling [29]. The characteristic of the pooling layer is reducing the complexity [28]. Pooling separates the image into several regions. A frequently used size for the pooling layer is 2x2. The difference between max pooling and L2 pooling is the operation applied on the given 2x2 pixels. Max returns an output of size 2x2 with the maximum activation values, while L2 returns a size 2x2 output with the square root.

Fully-connected layer

The fully connected layer contains neurons to connect two layers like other neural networks [28]. Each neuron is connected to both the input neuron and the output neuron. Moreover, this requires a lot of computation. Therefore, removing some of these connections is a common operation to do. Dropout is a technique that removes some of the connections through a random parameter [29].

Non-Linearity layer

Non-linearity layer limits the generated output or adjusts the values. In figure 2.1 the ReLU is the non-linearity layer. ReLU was introduced in the paper Hahnloser et al. [30], and is a fairly simple function. If output x is lower than 0 it returns 0 and if it is greater than 0 it outputs a linear

function. Returning a linear function means that the output value depends on the input value into the ReLU activation function [31]. The formula for ReLU is:

$$f(x) = max(0, x)$$

2.3.5 Transformers

[Explain what transformers are] Looking at the left block in figure 2.5,

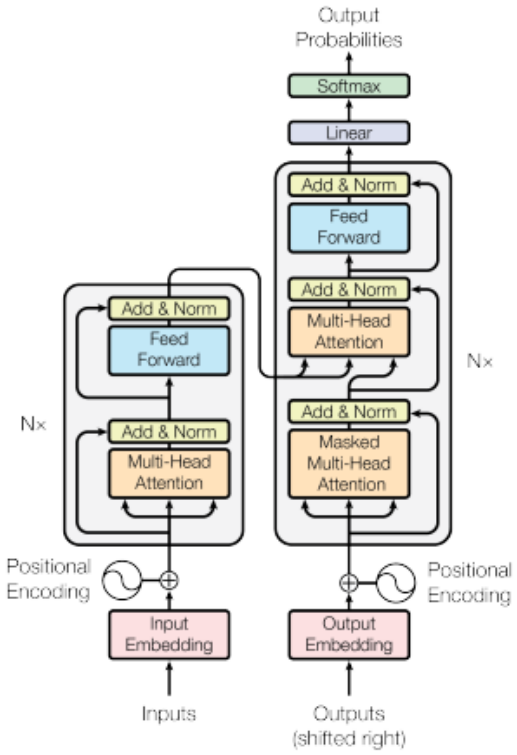

Figure 2.5: Transformer model architecture [32] (Left encoder, right decoder)

the input is first fed into an embedding layer. In this layer, the input is converted to vectors. Additionally, a positional encoding is appended to the input so the model understands the order of the input. This is done through a sin cosine function. Next, the encoder block contains the multi-head attention mechanism and a feed-forward network. The multi-head attention mechanism teaches the model connections between words.

The output from the multi-head attention block is fed into a feed-forward network. The objective of this is to train the network.

Looking at the right block in figure 2.5, the decoder unit contains similar layers as the encoder unit; two multi-headed attention layers and a feed-forward network. The first multi-headed attention layer acts similarly to the multi-headed attention layer in the encoder unit, except it masks words. The second multi-headed attention layer compares the input to the encoder block with the input to the decoder. This helps the model understand which words are more important. Last, the output from the second multi-headed attention layer goes through a feed-forward network, and then that output is run through a softmax function. This will generate a list of words with probabilities. The one with the highest probability will be the output.

In the next section, language models utilizing this architecture will be explained. These models will be benchmarked later in the thesis.

2.4 Language models

Today, several language models exist. Different models excel at various NLP tasks. The tasks vary from question and answering, text summarization, translation, labeling of categories, and more. In the earlier days of NLP, a basic rule-based approach was utilized. Now, the most prominent models are based on deep learning techniques, using different neural network architectures like CNN, RNN, and Long short-term memory [28], [33], [27]. In this section, different language models will be explained. These will be the ones benchmarked later in the thesis.

2.4.1 Embeddings from Language Models

Embeddings from Language Models (ELMo) is an improvement to the challenges machines have regarding understanding syntax, semantics, and polysemy, in natural language [34]. This is achieved by improving the skip-gram model. Moreover, the skip-gram model is an approach to predict

the context of a sentence, achieved when training the neural network with word pairs [35]. ELMo differentiates from the normal skip-gram model by assigning representation to each word, given the entire sentence [34]. Therefore, two words with identical spelling can produce different contexts. An experiment reported by Yang et al. [36] combined ELMo with RNN. The ELMo+RNN model scored the best accuracy with almost 1.5 percent higher accuracy, compared to other models. However, this experiment only had two classifications, positive and negative.

2.4.2 Bidirectional Encoder Representations from Transformers

Introduced by Devlin et al. [37], BERT is an abbreviation for Bidirectional Encoder Representations from Transformers. BERTs pre-training utilized a bidirectional approach, which means it is not limited to processing text from left to right. The language model randomly masks a couple of words in the input to achieve this. This is because BERT will look at the context of the whole input. Another approach utilized is next-sentence prediction. Next sentence prediction is a task that predicts or connects two sentences together, maintaining a long-term relationship between the sentences.

BERT is applicable for fine-tuning after the pre-training. In the paper [37], they show examples of a fine-tuned BERT model on the NLP tasks and labeled datasets; GLUE [38], SQuAD [39], and SWAG [40]. The BERT model is fine-tuned by adding a last layer that classifies the sentence. In the paper [41], the authors experimented with fine-tuning the BERT model upon the different tasks MNLI [42], SQuAD [39], and dependency parsing. The results across these tasks showed an accuracy of MLNI = 83.3, SQuAD = 89.2, and dependency parsing = 96.3

In the paper written by Acheampong, Nunoo-Mensah, and Chen [43] observations are made that BERT is the most explored transformer-based model for emotion detection in texts. Improvements have been made to the model, both for the sake of improvement and to apply the model to domain-specific areas. Examples of this are RoBERTa and DistilBERT.

2.4.3 DistilBERT

The objective of DistilBERT is to reduce the size of the BERT model and retain its language understanding [44]. DistilBERT utilizes knowledge distillation to compress the model. This aims to shorten the computational cost, which also reduces the environmental cost. The main difference in architecture is a reduction of layers. Across the same NLP tasks, DistilBERT showed a 3.9% reduction in accuracy on SQuAD and a 1.1% reduction on MLNI. Final results reported that distilBERT was 60% faster [44].

2.4.4 Robustly optimized BERT approach

Robustly optimized BERT approach (RoBERTa) is a proposed optimization of the model BERT, introduced in the paper by Liu et al. [45]. The optimizations for pre-training RoBERTa include; Removing the next sentence prediction loss, increasing the input to always be 512 tokens, larger batches of data, and applying masking several times on the same sentence. RoBERTa showed an improvement in the SQuAD tasks as well as MLNI.

2.4.5 BART

BART is a denoising autoencoder [46]. Denoising autoencoder means it accepts corrupted data and then reconstructs it. The architecture is similar to BERT, with the main difference being that the decoder layers perform cross-attention and remove the additional feed-forward network before predicting words. BART excels at text summarization as well as reconstructing corrupted documents.

2.4.6 Generative Pre-trained Transformer

Generative Pre-trained Transformer (GPT) model utilizes a semi-supervised machine learning [47], which means unsupervised pre-training and supervised fine-tuning. This is done by pre-training on unlabeled data and fine-tuning the model on labeled data.

GPT is based on the transformer architecture, with 12 transformer layers. The optimizer is Adam [48], and the activation function is a Gaussian error linear unit (GELU) [49]. Fine-tuning GPT consists of manipulating the input. For example, for text entailment, the input is "premise + hypothesis." GPT can perform several tasks: text classification, similarity in texts, Q&A, and test generation. When released, it achieved state-of-the-art results for the task, MLNI. For sentiment analysis, it scored 91,3% accuracy on The Stanford Sentiment Treebank (SST-2) task.

Generative Pre-trained Transformer-2

In later years, the research laboratory OpenAI team improved the GPT model. The GPT-2 model by Radford et al. [50] was primarily improved by increasing both; the number of parameters and the size of the dataset. GPT-2 received state-of-the-art results in 7 out of 8 language modeling tasks using a zero-shot approach. Zero-shot is a type of machine learning where the predicted class was not introduced during training. [51] The different tasks were summarization, comprehension, translation, and Q&A.

Generative Pre-trained Transformer-3

On June 11th 2020, beta version of GPT-3 was released by the OpenAI research laboratory [52]. The significant difference between GPT-3 and GPT-2 are the following; the size of parameters is 175 billion compared to GPT-2's 1.5 billion, the number of layers increased, and it is trained on a corpus of 499 billion tokens of data scraped from the web [53]. Results from the experiments conducted by the OpenAI team showed limitations of the GPT-3 model, including synthesizing text, repeating itself semantically on a document level, and understanding specific "common sense physics" sentences [52]. An experiment by Del Arco et al. [54] compared supervised and zero-shot learning. The experiment showed that GPT-3 is a promising option for emotion classification when no labeled dataset is available.

2.4.7 Huggingface

Huggingface is an open-source project which aims to make language models available for everyone. Moreover, it contains several datasets and models. The transformer library, a Python package, contains open-source implementations of the transformer language model. This thesis utilizes the BART model [2] and the DistillroBERTa model created by Jochen Hartmann[55] for sentiment analysis which is found on the Huggingface website [3].

2.5 Datasets

Datasets are useful information gathered that language models can train on and be tested on. Many datasets already exist, which can be applied to different types of language models. Some notable sentiment-based datasets are MELD [1], IMBD dataset[2], and more.

Multimodal EmotionLines Dataset

MELD (Multimodal EmotionLines Dataset) is a dataset for the task of emotion recognition in conversations [56]. MELD is an extension of the existing EmotionLines dataset. There are 13,000 utterances from the TV show Friends. The labeled emotions are annotated by graduate students. Moreover, they could choose from seven emotions, anger, disgust, fear, joy, neutral, sadness, and surprise.

2.5.1 Labeled dataset

In the paper by Lammerse et al. [57], a user study was conducted. This user study aimed to annotate emotions in certain portions of a transcript. The participants were of different age, education, and gender. This resulted in a dataset with children's utterances labeled with emotions. The emotions are

[2]https://huggingface.co/facebook/bart-large-mnli
[3]https://huggingface.co/

annotated with several participants' opinions. An emotion was selected if it had 20% more votes than the emotions with the second most votes.

Introduced in section 3.1.1, this dataset contains sentences with emotions annotated through a user study. The reason for using a dataset with annotated data is to explore the language model's ability to predict children's emotions. Hence, for this thesis, we decided to use a labeled dataset that was created in the paper by Lammerse et al. [57]. The dataset was annotated in a survey containing 21 participants. The gender distribution among them was 52.4% female and 47.6% male, with ages ranging from 26 to 65. Additionally, the dataset has 44 rows with the number of sentences ranging from 1 to 7. The last seven rows in the dataset contained single utterances, with the rows being emotion-labeled based on the previous utterances. Each row was labeled with one of these four emotions: sadness, anger, fear, and enjoyment.

2.5.2 Abusive transcripts dataset

Centre for Investigative Interviewing[4] is a research and training center for improving investigative interview practices. The team consists of several different fields varying from psychology, criminology, education, and more. Professionals conducted interviews following specific guidelines, which were compiled into several transcripts. These transcripts represent the dataset for abusive behavior.

2.5.3 Childes non-abusive transcripts dataset

Brian MacWhinney organized a project called TalkBank [5]. It is a project with several repositories for spoken language. In this thesis, the component Childes [58] was scraped into a non-abusive transcript database.

Childes is a corpus that contains dialogues between a child and an interviewer, sometimes also the parent is present [58]. There are several

[4]https://www.investigativecentre.com/

[5]https://talkbank.org

different subdirectories in the Childes corpus. The English Fletcher Corpus [59] has 72 interviews between a female adult and a child. The ages of the children interviewed are 3, 5, and 7. The aim of this project is divided into three categories. First, was creating a database. Second, identify important grammatical and lexical features. Last, use this information to help language-impaired children.

2.6 Local Interpretable Model-agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) is a Python library used to explain the prediction of a classifier [8]. Using the explainer module provided by LIME we can get a visual representation of which input sections provided the predicted result. For example, in the abusive transcripts, some sections contain many more abusive descriptions than others. With LIME we can extract these sections and insert them into a heatmap, hopefully gaining a representation of important parts of the transcripts.

2.7 Related work

2.7.1 Emotional recognition in conversations

Emotional Chatting Machine (ECM) [60] is a proposed solution for expressing emotion and feeling. The ECM is trained with supervised learning using a manually annotated dataset. Designed with a sequence-to-sequence generation model with relevant changes related to correctly categorizing emotions, an internal state balancing emotions and grammar, and an external state to generate defined emotional expressions.

ECM is designed with a sequence-to-sequence model. Their model uses the encoder-decoder framework with GRU units which are recurring gated units. GRU's are an improvement on a long short-term memory neural network [61]. The framework works by having the encoder convert an

input text to vector representation at each time unit. Then the decoder takes the decoded word from the last time unit and an input vector to update its state. When this state is updated, an output is generated based on the probability computed by the state of the decoder [60].

A limitation of the ECM is that the response is dependent on the relation between the responder and poster. Meaning a friend might answer with a different emotion compared to a total stranger. This could be solved with either respond with a more generalizing answer or give the ECM a personality [60].

Contrastive and Generation-enhanced BART (CoG-BART) is a language model for classifying emotions for sentences in a text [62]. The architecture consists of a dialogue-level transformer, SCL loss function, text generation for better understanding contexts, and a pre-trained BART model. The model was tested on sentiment-labeled datasets. It achieved improved results compared to BERT, RoBERTa, and BART when classifying emotions for MELD, EmoryNLP, IEMOCAP, and DailyDialog datasets. However, it is worth noting that BERT, RoBERTa, and BART are not approaches for classifying emotions.

HiTrans is a transformer-based model introduced in the paper by Li et al. [63]. HiTrans architecture consists of a low-level transformer model based on BERT, which feeds the output into another transformer [63]. This achieves long-range context information over a conversation. The output of the last model is inserted into a multi-layer perceptron (MLP), then classified with a biaffine classifier. The MLP determines the emotion, while the biaffine classifier establishes which speaker is uttering. The results across the MELD, EMoryNLP, and IEMOCAP datasets were slightly lower than CoG-BART.

2.7.2 Explainablility

In the paper, by Kapoor and Kumar [64], the authors propose to extract important features from spectrogram images using CNNs in the context of speech emotion recognition. Moreover, use these features to detect stress and anger. They argue that detecting anger and stress early on is important because they can impact a person's mental and physical health.

In the paper, by Mustaqeem, Sajjad, and Kwon [65], the authors propose using speech emotion recognition to convert sequences of speech into spectrograms. These spectrograms are then passed into a CNN for analysis. The sequences of speech used are the key segments instead of whole sentences. The goal is to detect the final emotional state.

The paper, by Pham et al. [66], proposes extracting a set of features using overlapping sliding window technique for speech emotion recognition. Additionally, they used a DNN to classify the emotional state of the extracted features.

2.8 Summary

The design and creation of a virtual-based avatar require several different components, audio, visual, and textual components. In this thesis, we will focus on the textual part of the avatar, more specifically the sentiment component. In the paper by Vaswani et al. [32], an encoder-decoder architecture was introduced. Almost all the models examined in this chapter are transformer-based architecture. Many of these models performed well on different NLP tasks. These results were studied and considered to decide which models to benchmark in the next chapter.

Related work proposes using speech represented as spectrograms for extracting emotional states. However, using audio would impose privacy concerns, as the scope of the thesis concerns abused children. Instead, we use textual utterances from interviews with children displayed as line

charts as emotional profiles.

In the next chapter, we will look at all the steps for generating heat maps of the extracted features. First, data from the Fletcher corpus needs to be converted into a usable dataset. Then, language models need to be benchmarked to get an overview of which models should be included in later experiments. With the selected models, a group of line charts will be created. These line charts will represent the images of the emotional state of a child which we insert into a CNN. We then extract important features from the CNNs prediction using the LIME [8] Python package and display them in a heat map.

Chapter 3

Design and implementation

There are several language models applicable to sentiment analysis. However, the scope of this thesis is limited to children's emotions so we benchmarked the language models to gain information about their capability to predict children's emotions.

In this chapter, we discuss the steps executed for explainability analysis. Starting with discussing the datasets utilized for both benchmarking and the creation of sentiment profiles. Then, we introduce the benchmarking process, where the ability of the language models' are tested within the scope of the thesis. Using the capable models we explain how they were applied to creating sentiment profiles. Finally, we explain the generating of heat maps used for explainability analysis.

3.1 Datasets

The two tasks of benchmarking and explainability analysis required datasets containing transcripts between a child and an interviewer. The dataset utilized for benchmarking had to contain annotated emotions so we could asses the language model's abilities. The two datasets used for explainability analysis required abusive and non-abusive transcripts. As these were not available, we had to gather them into a dataset on our own.

In this section, we will go into more detail about collecting these datasets, as well as provide more information about them.

3.1.1 Benchmarking dataset

Introduced in section 3.1.1, this dataset contained sentences with emotions annotated through a user study. The reason we used this dataset was to explore the different language model's ability to predict children's emotions. In detail, the dataset had 44 rows, containing a number of sentences ranging from one to seven. The format of the dataset was:

$$ID, Utterances, Number\ of\ sentences, Sentiment$$

Excluding the last seven rows, the dataset contained several sentences from abusive transcripts. Each row was annotated with one of these emotions: sadness, anger, fear, or enjoyment.

3.1.2 Abusive dataset

The abusive dataset contained 20 transcripts obtained from the Centre for Investigative Interviewing[1]. These transcripts documented abusive investigative interviewing between professionals and children, as explained in more detail in section 2.5.2.

3.1.3 Non-abusive dataset

The Talkbank project[2] contains several databases of interviews with children. In this thesis, we utilized the database of Fletcher [59], for more information about this project see section 2.5.3. From this project, we scraped the transcript from interviews with children of the age of seven. We chose to use these transcripts as they contained more structured and worded utterances compared to younger children. However, the transcripts consisted of several typographical symbols and words such as "xxx", "[] with words inside", and "& as and". To ensure that the model

[1]https://www.investigativecentre.com/

[2]https://talkbank.org

interpreted the sentences correctly we removed these. Additionally, there were several cases where one person had multiple utterances in a row, to solve this we concatenated them into a single utterance until the other person spoke. The resulting data was then turned into a transcript for the abusive dataset. Moreover, the format of the transcript was

Transcript ID, Interviewer utterance, Child utterance, age

. In total, we scraped 20 non-abusive interviews into non-abusive transcripts. Hence, matching the total number of abusive transcripts.

3.2 Benchmarking

The aim of benchmarking the language models was to identify the best-suited model for sentiment analysis on children's utterances. We conducted the benchmarking process using the dataset mentioned in section 3.1.1. Our main objective during the benchmarking process was to see which model predicted the most correct sentiment. However, for the explainability analysis part of this thesis, we required sentiment profiles consisting of confidence scores predicted by the language models. Therefore, we did an additional analysis of the language model's ability to return confidence scores of other emotions. In total, We benchmarked five models: DistilRoBERTa [55], bart-large-mnli (BART) [46], GPT-3 [52], HiTrans [63], and CoGBart [62].

3.2.1 DistilRoBERTa

The DistilRoBERTa [55] language model was found on the website Huggingface[3]. It was trained on six different emotion-labeled datasets; Crowdflower[4], Emotion Dataset [67], GoEmotions [67], ISEAR, Vikash (2018), MELD [56], SemEval-2018 [68]. Moreover, the model predicts

[3]Huggingface is an open-source project which aims to make language models available for everyone. https://huggingface.co

[4]Crowdsourced dataset, https://huggingface.co/datasets/tasksource/crowdflower

Ekman's [69] six basic emotions, plus a neutral class. The results achieved with this model on the emotion-annotated dataset was

$$Accuracy : 21/44$$

$$Sadness : 0, Fear : 11, Anger : 3, Enjoyment : 7$$

3.2.2 BART

As described in section 2.4.5, BART is a denoising autoencoder. We used the bart-large-mlni[5] model for benchmarking, which is available on the Huggingface platform. This model was trained on the MultiNLI dataset [70], which consists of 433k sentence pairs with various genres, such as government, telephone, fiction, and more [46]. BART was benchmarked with the emotions of sadness, fear, anger, and enjoyment. The results achieved with this model on the emotion-annotated dataset was

$$Accuracy : 21/44$$

$$Sadness : 0, Fear : 8, Anger : 3, Enjoyment : 10$$

3.2.3 Applying GPT-3 model to predict sentiments

We used the text-davinci-003 GPT-3 model by openAI[6] [52]. It was designed to follow instructions and perform tasks based on prompts given as input. GPT-3 was sensitive in what it returned based on the prompt given. Therefore, we had to design a detailed prompt with specific instructions for achieving the best possible predictions. The sentiments used in the benchmarking were sadness, anger, fear, and joy. For a single sentence, we used the prompt seen in figure 3.1. When combining sentences into a "story" we used the prompt seen in figure 3.2. The results achieved with this model on the emotion-annotated dataset was

$$Accuracy : 29/44$$

$$Sadness : 0, Fear : 17, Anger : 4, Enjoyment : 6$$

[5]https://huggingface.co/facebook/bart-large-mnli

[6]https://openai.com

```
prompt = "Use sentiment analysis and classify the text using these sentiments: "\
            + sentiments + row[1]
```

Figure 3.1: Single sentence input GPT-3

```
prompt = "Use sentiment analysis and classify the text using these sentiments "\
            + sentiments + ". " + story
```

Figure 3.2: Story input GPT-3

3.2.4 Applying other models to predict sentiments

In this section, we review the benchmarking of the CoG-BART [62] and the HiTrans [63] models. We utilized the emotion-labeled dataset discussed in section 3.1.1, for the benchmarking process.

CoG-BART contained four pre-trained models available for use. We used the pre-trained model that was trained on the MELD [56] dataset. The results achieved with CoG-Bart were not as good as the results achieved with GPT-3, DistilRoBERTa, and BART. A reason for this could be that we used it wrong, because CoG-BART used a different setup for prediction on datasets, making it challenging to apply it to the emotion-labeled dataset.

HiTrans did not include a pre-trained model so we had to train it first. There was no available dataset to be used for training this model to learn children's emotions. Therefore, we trained it using the MELD [56] dataset as it was easily accessible. The results achieved with HiTrans were similar to CoG-BARTs. Similar as with CoG-BART, this could be our own fault. Training the model depended on earlier versions of PyTorch, and the computer we used had a GPU that did not support this. Therefore, we had to make some changes to the model, which could have affected the results.

3.2.5 Model selection

We based the selection of models on two factors, their accuracy score and their ability to return confidence values for each emotion. In the table 3.1

the scores from the most prominent models are presented.

Models	Accuracy	Sadness	Fear	Anger	Enjoyment
DistillRoBERTa	21/44	0/1	11/24	3/7	7/12
BART	21/44	0/1	8/24	3/7	10/12
GPT-3	27/44	0/1	17/24	4/7	6/12

Table 3.1: Benchmarking scores for the different language models

Accuracy score

The best accuracy score achieved was with the GPT-3 model, with a score of 29/44. In comparison, DistilRoBERTa and BART achieved an accuracy score of only 21/44. However, DistilRoBERTa had a disadvantage compared to the other models in that it did not provide us with a choice of selecting emotions to include. If the annotations of the benchmarking dataset consisted of the same emotions as the ones used by DistilRoBERTa we might have seen a higher accuracy for this model.

An issue in the emotion-labeled dataset that skewed the true value of the accuracy score was that a few of the rows of sentences was incomprehensible without the context. Examples of this are displayed in table 3.2. For the first sentence in table 3.2, the annotated emotion was fear. DistilRoBERTa, BART, and GPT-3 all predicted the emotion of sadness for this row in the benchmarking. Without the full context, this appeared to be the appropriate sentiment. However, the context of the interview is that the interviewer asked, "Okay, and then what happened when he hit you a hundred times?". Arguably, this could indicate that the child would express the emotion of fear. The second sentence appears fairly neutral upon inspection. The context of the conversation about the child's dad's last name and when the child last saw his/her dad. The models all predicted sadness, the correct sentiment was enjoyment.

ID:	Sentence:
1	Crying. Then he left. He left.
2	it's Brown. About two weeks ago. No.

Table 3.2: Two rows from the labeled dataset

Confidence score of the sentiments

For explainability analysis, we used the confidence scores predicted by the language models to create a sentiment profile of the child. Therefore, it was important that the models returned a usable confidence score for all the emotions. DistilRoBERTa and BART both predicted confidence scores suitable for creating a sentiment profile. On the contrary, the confidence score returned by the GPT-3 model was often 0.0 for other emotions than the one predicted, causing the sentiment profile to be difficult to use in explainability analysis. Therefore, the selected models were BART and DistilRoBERTa, although they had a lower accuracy for predicting correct sentiment.

3.3 Sentiment profile

As mentioned earlier, sentiment profiles are confidence scores appended together to show the emotional state of a child throughout a transcript. For explainability analysis, we will use these to detect important features behind a DNN's prediction. Several design choices had to be made during the creation of sentiment profiles, including the choice of emotions, parameter values, and length. In this section, we will present our reasoning for the decisions we made.

3.3.1 Emotions included as prediction

There were several reasons for including neutral when creating the sentiment profiles. Firstly, according to the paper Gasper, Spencer, and Hu [71], neutral affection is an important sentiment that provides information

about a person's emotional state. Secondly, excluding neutral resulted in an overestimation of the value of other emotions. Therefore, we argue that including the emotion of neutral improves the quality of the sentiment profile. For instance, when we applied the BART language model to the sentence "Hello, my name is John.". We see that table 3.3, depicts that BART predicted a higher confidence score for the emotions compared to table 3.4. Further, inspection displayed that the confidence values predicted doubled. Resulting in seemingly neutral utterances being labeled with a much higher confidence score than reasonable.

Joy	Anger	Fear	Disgust	Sadness
0.34	0.20	0.19	0.17	0.10

Table 3.3: BART's prediction without neutral included

Neutral	Joy	Anger	Fear	Disgust	Sadness
0.51	0.16	0.10	0.10	0.09	0.05

Table 3.4: BART's prediction with neutral sentiment included

Seven emotions were initially selected to represent the sentiment profile. These were Ekman's [69] six basic emotions of anger, disgust, fear, joy, sadness, and surprise plus neutral. These ones were pre-selected in the DistilRoBERTa language model. Similarly, the BART language model included the same emotions, except for surprise. We decided to remove surprise as it was usually overestimated. For instance, when we applied the BART language model to the sentence "Hello, my name is John.", table 3.5 reveals a high confidence score for the emotion of surprise, even though the sentence appears fairly neutral. Therefore, we excluded surprise as an emotion for the BART language model.

3.3.2 Variable parameters

For better predictions of emotions, we introduced two variables, window size, and threshold. Window size variable controlled the full context in-

Surprise	Neutral	Joy	Anger	Fear	Disgust	Sadness
0.37	0.27	0.14	0.07	0.06	0.05	0.03

Table 3.5: BART's prediction with surprise included

cluded for a prediction. Moreover, the value of the variable decided how many previous sentences to include in a prediction. For example, a made-up transcript consisting of three sentences, depicted in table 3.6. In figure 3.3, the BART model only predicted the sentence "no". Moreover, BART assigned joy with a confidence score of 0.02. However, when inspecting the full context of this made-up transcript a reasonable deduction is that the child is playing and enjoying themselves. When including the full context of the transcript as seen in figure3.4, we see a confidence value of 0.93 for the emotion joy. This improvement is significant as single words were commonly used in the transcripts. Incorporating the variable window size resulted in greater accuracy for the language models. We decided to use a window size of 5 when creating the sentiment profiles for both language models

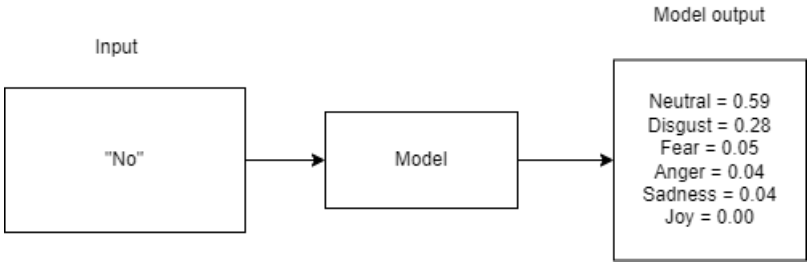

Figure 3.3: Single sentence from a dialog

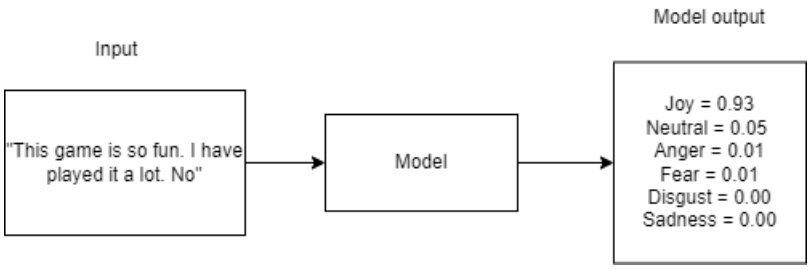

Figure 3.4: Dialog with the context

Speaker:	Sentence:
1	This game is so fun
1	I have played it a lot
1	No

Table 3.6: Example interview transcript

An additional variable called threshold was introduced in order for the language models to identify sentences that could impose sudden changes in the emotional state. We utilized this variable by applying the language models to every sentence alone. If the confidence score of an emotion exceeded the threshold value, the language models ignored the context. Studying the results from figure 3.5 and figure 3.6 shows that the emotion of joy is drastically changed when the full context is included. Specifically, in figure 3.6, the value of joy was 0.02, while in figure 3.5, it increased to a score of 0.90. As the transcripts contained sections where the topic changed, potentially altering the emotional state. We argue that by including the threshold variable, the language models could predict these changes accordingly. We decided to use a value of 0.8 for threshold when creating the sentiment profiles for both language models

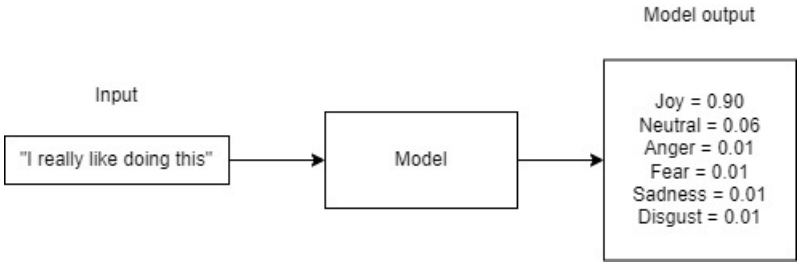

Figure 3.5: Single utterance from a dialog

The language models DistilRoBERTa and BART only used the child's utterances as input when predicting the confidence values. This was done to prevent incorrect predictions caused by including the interviewer's utterances. Moreover, the language models could predict a combination of

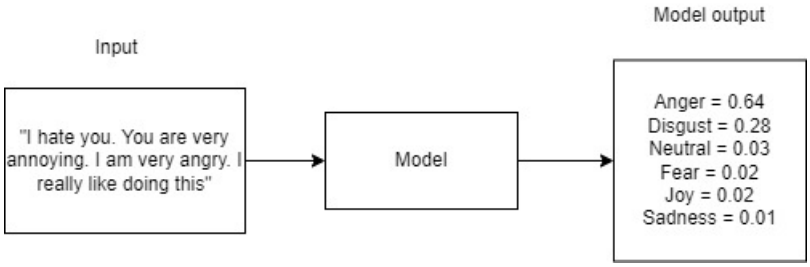

Figure 3.6: Context included from a dialog

the interviewer's and the child's emotions. An example of this is shown in figure 3.7, the child says "I am not afraid at all" while the interviewer says "I am very afraid". The BART language model predicts a confidence score of 0.87 for fear. As a result, we decided to exclude the interviewer's utterances. This could cause the context of a transcript to be lost, but this was deemed as less important than distorting the sentiment profiles with the interviewer's emotions.

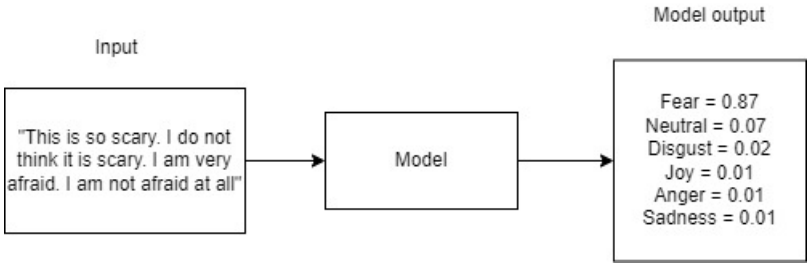

Figure 3.7: Prediction with both speaker's utterances included

3.4 Implementation of generating heat map

To generate the heat maps we first had to train a DNN. We decided to use a CNN as it excels at analyzing pictures, and the sentiment profiles can be viewed as an image of an emotional profile. We used two different types of sentiment profiles, abusive and non-abusive. For the CNN to analyze the sentiment profiles we had to convert them into readable data. Each sentiment profile was converted into the shape

$$transcript_i = [[Anger], [Sadness], [Fear], [Disgust], [Joy]].$$

Additionally, because the size of the transcripts differed, we either truncated or extended each separate emotion by repeating the confidence scores to a padding variable. Moreover, if the padding is 200, each sentiment profile input would contain 200 confidence values for each emotion, 200 confidence scores for anger, 200 confidence scores for sadness, etc.

The implementation of the CNN consisted of a single convolution layer (line 4 in figure 3.8), a pooling layer (line 5 in figure 3.8), and two fully connected layers (line 6 and 7 in figure 3.8). The convolution layer was defined with input channel = 1, output channels = 16, and kernel size = 3. The input was first run through the convolution layer in the forward phase. Then ReLu activation operation was performed on the output of the convolution layer and run through a max-pooling layer. The pooling layer was set up with a filter size = 2x2 and stride = 2. We then reshaped the tensor object so it could be input into the fully connected layers. Before finishing the forward phase, the input was run through the last fully connected layer. Additionally, the optimizer used was the Adam optimizer with a learning rate of 0.001 and an epsilon value of 0.1. In short, the Adam optimizer is a stochastic optimization algorithm [48]. We used the Epsilon hyperparameter to avoid that division with zero altered the results. Additionally, the cross-entropy loss function was used. The prediction classes for the CNN were 0 for non-abusive and 1 for abusive.

```
1  class SimpleCNN(nn.Module):
2      def __init__(self):
3          super(SimpleCNN, self).__init__()
4          self.conv1 = nn.Conv2d(1, 16, 3, padding=1)
5          self.pool = nn.MaxPool2d(2, 2)
6          self.fc1 = nn.Linear(16 * 8 * 25, 64)
7          self.fc2 = nn.Linear(64, 2)
8
9      def forward(self, x):
10         x = self.pool(F.relu(self.conv1(x)))
11         x = x.view(-1, 16 * 8 * 25)
12         x = F.relu(self.fc1(x))
13         x = self.fc2(x)
14         return x
```

Figure 3.8: Simple Convolution Neural Network

The CNN was trained and tested with five iterations using k-fold cross-validation. K-fold cross-validation was used because of the small sample size in the datasets. With each iteration, the abusive and non-abusive datasets were split into training and test sets. We usually saw a prediction rate of 100% accuracy. Especially, when we introduced the epsilon value in the Adam optimizer.

After the CNN was trained and tested, we extracted the important features for the prediction. This was done by using LIME [8]. LIME returned an array of all the important features with a value corresponding to the importance of the prediction. We then inserted these features into a heat map for easier analysis and visualization. Moreover, this enabled us to easier locate certain patterns and groups of features. Each feature corresponds to the confidence score by a model for a transcript. For example, feature 1 represents the confidence score predicted for anger on the sentence at index 1 in the transcript. Additionally, feature 201 corresponds to the confidence score predicted for sadness on the sentence at index 1. This correlates to the padding variable used by using padding as a modulus on the representing number of the feature extracted.

3.5 Summary

In this chapter, we went into detail about the setup for generating the heat maps used in explanation analysis. The process started with gathering and creating two datasets, abusive and non-abusive. We gathered these as we could not find available datasets to use within the scope of the thesis. Additionally, we were only able to gather a total of 20 abusive transcripts.

Then, we benchmarked different language models on an emotion-labeled dataset to determine their ability to predict children's emotions. Choosing the correct model included inspecting their accuracy score and the ability to return confidence scores of all the sentiments. GPT-3 achieved the best accuracy score. However, GPT-3's ability to return confidence scores for

other emotions besides the one it chose was limited. Additionally, it was shared across the Avatar project and it would require a lot of time to label 40 transcripts. BART and DistilRoBERTa had similar accuracy to GPT-3 model. Therefore, these models were chosen for creating the sentiment profile used as input for the DNN. For creating the sentiment scores we used two variables to further increase their performance. The two variables were threshold, to control emotional changes within a transcript, and window size, to include context to the prediction. We used values of threshold = 0.8 and window size = 5.

Lastly, we implemented a neural network to analyze the sentiment profiles. We utilized a CNN to classify the sentiment profiles of the abusive and the non-abusive transcripts. By doing this we were able to extract the important features in a sentiment profile. We plotted the features onto a heat map which we will analyze in the next chapter. [7]

[7] https://github.com/knutsiv/Master

Chapter 4

Results

As discussed in the previous chapter, we utilized a CNN for analyzing the image of a child's emotional state throughout a transcript. Moreover, the CNN predicted whether the transcript was abusive or non-abusive. By using explainable AI we can extract the important features behind the prediction. In this chapter, we will inspect the heat maps created to represent these important features. Moreover, we will mainly observe and examine, then draw conclusions in the next chapter.

4.1 Metrics

First, we start by introducing the heat maps used for analysis. Furthermore, we will discuss the extracted features and their correlation to a transcript. Finally, we argue against including neutral as an input emotion for CNN.

4.1.1 Heat maps

For explainability analysis, heat maps served as the primary metric. The x-axis of a heat map represents the location of every possible feature in the CNN, while the y-axis represents the emotions. Moreover, each feature corresponds to the index in the input sentiment profile. To illustrate this, we will examine an example heat map.

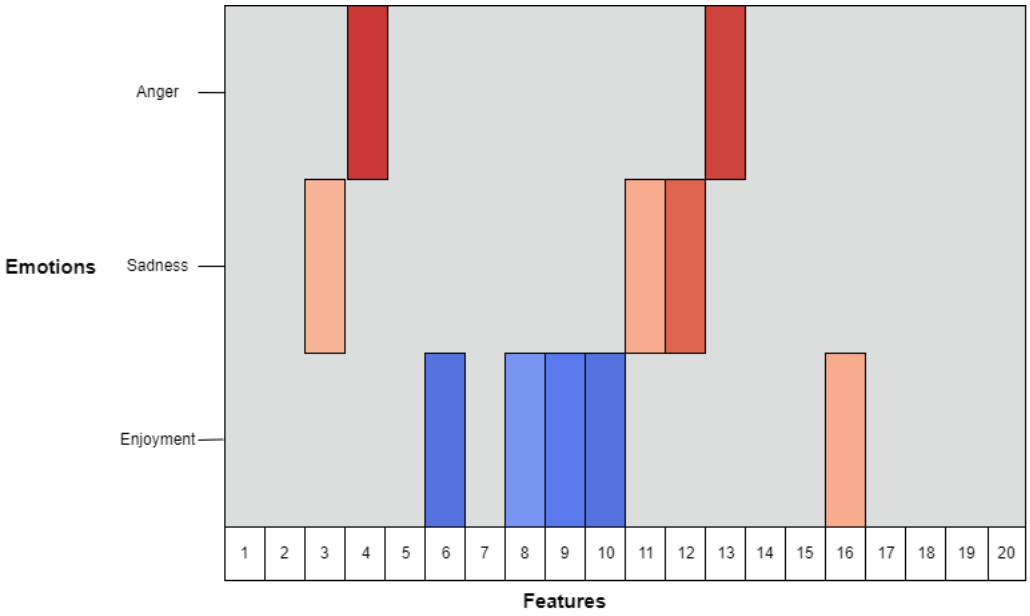

Figure 4.1: Example heat map

In the heat map in figure 4.1 the sentiment profile analyzed consists of 20 confidence values for each of the three emotions, anger, sadness, and fear. Moreover, the transcript to which the sentiment profile relates consists of 20 sentences. When we discuss important features we will refer to their location on the x-axis and then their location on the y-axis. For example, the top left feature in the heat map in figure 4.1 is referred to as feature 4 for anger. This feature could be translated into the predicted confidence score for anger on sentence 4 in the sentiment profile.

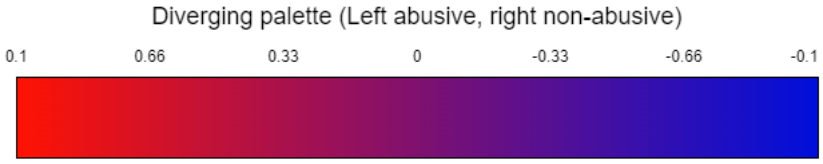

Figure 4.2: Diverging palette indicating the strength of a feature

In the example heat map in figure 4.1, we see a total of 10 red and blue boxes plotted onto the heat map. This means we instructed LIME to extract the ten most important features for the prediction. Additionally, there are two colors present in the different boxes plotted, red and blue. This is which

type of prediction the feature contributed towards. Extracting features with LIME assigns an importance value to every index in the input. Figure 4.2, shows how the intensity of the two colors corresponds to the importance value. Analyzing the example heat map in figure 4.1, we can identify that features 6-10 for enjoyment contribute to a non-abusive prediction, while features 4 and 13 for anger contribute to an abusive prediction.

4.1.2 Sentiments included

Initially, the sentiment profiles created included emotion neutral. This design choice was described in section 3.3.1. In this section, we will discuss and analyze the reason for excluding neutral emotion when analyzing sentiment profiles. The sentiment profile contained confidence scores for the emotions anger, sadness, fear, disgust, neutral, and joy. However, based on the heat maps generated with neutral included, it was observed that the neutral emotion significantly influenced the DNN's decision. The observations showed that especially the DistilRoBERTa heat maps were impacted, for both abusive and non-abusive sentiment profiles. For BART, however, it was mainly the non-abusive heat maps that were affected.

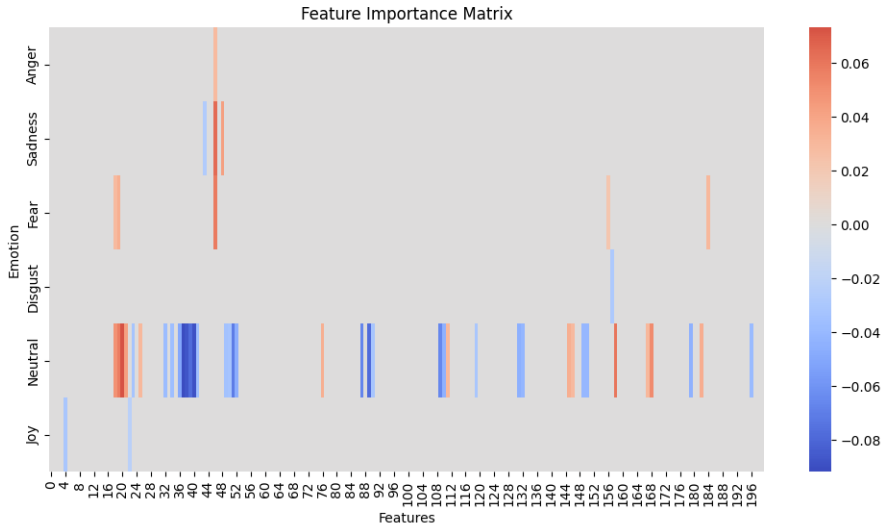

Figure 4.3: Heat maps with neutral included DistilRoBERTa non-abusive transcript 15

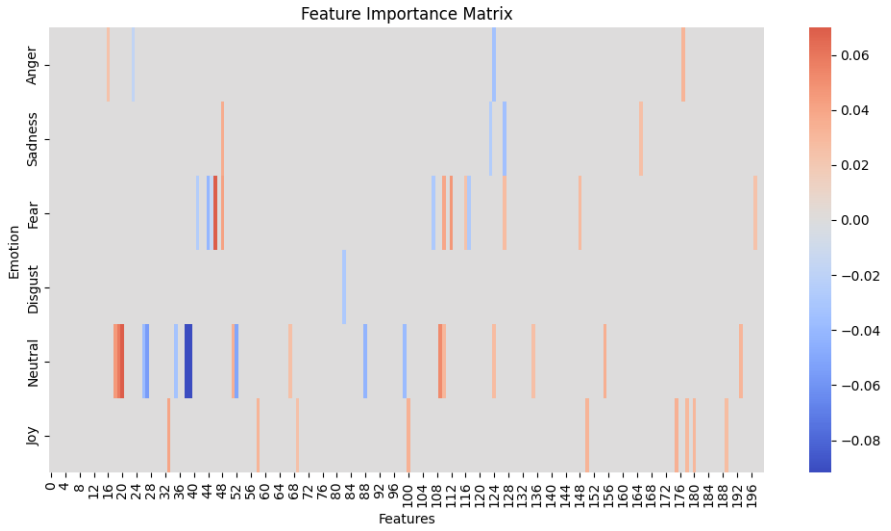

Figure 4.4: Heat maps with neutral included BART non-abusive transcript 18

Type	Transcript ID	Model	Nr of neutral features
Non-abuse	18	BART	8
Abuse	4	BART	5
Non-abuse	15	DistilRoBERTa	9
Abuse	13	DistilRoBERTa	10

Table 4.1: Number of neutral importance values out of top 10

When examining the heat map in figure 4.3 and the heat map in figure 4.4, it is clear that the majority of the important features extracted lie in the neutral row. Table 4.1, displays how many of the top 10 features with the highest importance value were neutral. With neutral included, other emotions appear insignificant for the DNN's prediction. In the context of sentiment analysis, this thesis aims to gain insight into the black box that neural networks represent. We want to solve the issues with transparency and interpretability of DNN's predictions. To reach this objective, it was decided to exclude neutral from the explainability analysis as the extracted features weighted this emotion too much. Another aim of this thesis is to

extract sentences based on their importance in a prediction. Analyzing the sentences extracted was more interesting when it was labeled as another emotion than neutral.

4.2 Results from abusive and non-abusive transcripts

For both datasets, we have 20 CSV files of sentiment profiles from both models, BART and DistilRoBERTa. Five emotions were included, anger, sadness, fear, disgust, and joy. The padding was set to 200 meaning each sentiment profile contains 200 confidence scores for each emotion, totaling 1000 confidence scores.

4.2.1 Results from the abusive transcripts

In this section, we will analyze the heat maps of the important features extracted, for the abuse prediction. Therefore, the main focus will be on the red features extracted.

BART transcript 1

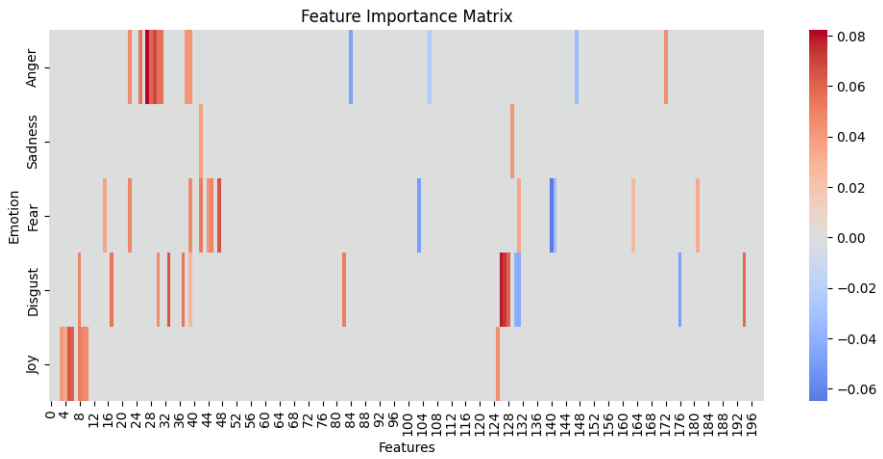

Figure 4.5: Heat map for BART prediction transcript 1 from the abusive dataset.

Comparing the heat map in figure 4.5 with the line chart in figure 4.6 reveals a correlation between the two. The line chart in figure 4.6, has

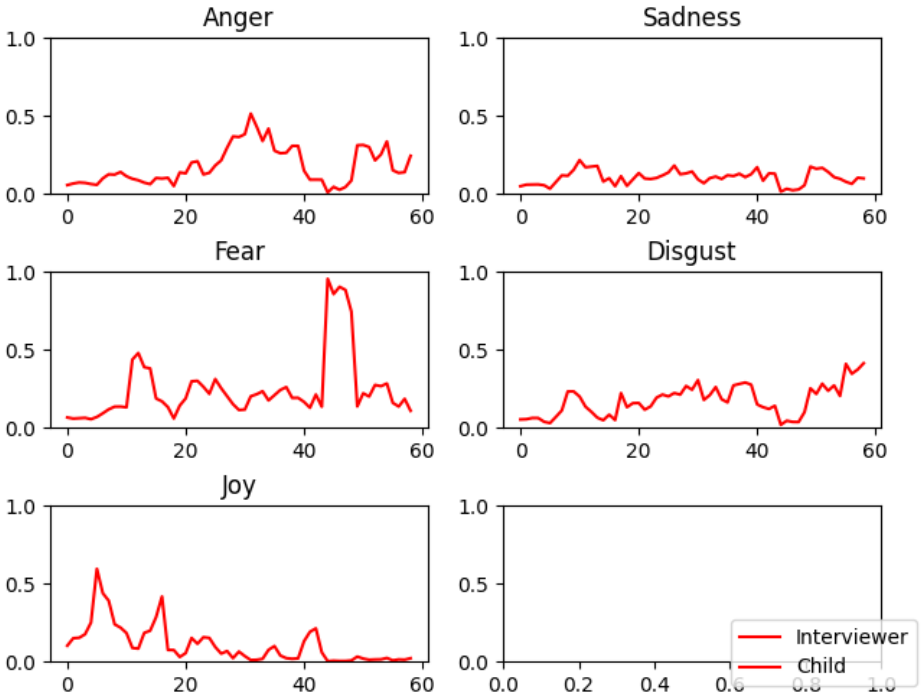

Figure 4.6: Line chart for BART prediction transcript 1 from the abusive dataset.

various local maxima's for different emotions that can be linked to groups of important abusive features extracted in the heat map in figure 4.5. This is particularly noticeable for the emotions of anger, fear, and joy.

In the heat map in figure 4.5 we can locate groups of important features extracted between the ranges of 25-32 for anger, between the ranges of 42-47 for fear, and between the ranges of 5-10 for joy. This indicates that the CNNs prediction for this transcript was based on these groups of features.

The first transcript in the abusive dataset starts with mentioning a fun activity that the child recently engaged in, causing a noticeable rise in joy at the beginning of the line chart in figure 4.6. However, around index 25 in the transcript, the conversation shifts to the abuse the child suffered. This is reflected in the important features between the ranges of 25-32 for anger. The anger features translated into sentences are displayed in table 4.2. The features between the ranges of 42-47 for fear translated into sentences, discuss the person who carried out the abusive actions. Moreover, the child

utters, "No because I'm afraid of Mark" and then elaborates the reasons why.

Feature	Sentiment	Sentence
25	Anger	Well they're red and they sting and they hurt when my shirt is touching them
26	Anger	Because I got a punishment
27	Anger	Well I left a mess in the kitchen and then Mark said that's enough and you really do need to get a punishment
28	Anger	Well I got some crackers with cheese from the fridge and I left cracker crumbs on the table and then Mark said get upstairs
29	Anger	Well And then I went upstairs and he said go up to your room and then he stomped outside and then he came upstairs
30	Anger	And then he told me that I should take off my dress and sit in the chair and then I got the punishment
31	Anger	He took the stick and he just kept hitting it on my back

Table 4.2: Table with sentences from transcript 1 from the abusive dataset.

BART transcript 4

There are two interesting groups of important abusive features observed in the heat map in figure 4.7. Analysis of the line chart in figure 4.8, displays that these groups correspond to local maxima for the relevant emotion, similar to what we saw in the previous heat map in figure 4.6. Specifically, features 3-14 for joy and features 19 and 20 for disgust. Although features 3-14 for joy represent a substantial number of sentences, we can summarize that the child discusses bathing with someone, which starts off innocently

but turns abusive around index 19. The sentences in table 4.3 are indexes 19 and 20. This is where the abusive behavior is described in the transcript. It is important to note that the abusive behavior is only described in this section of the transcript. The rest of the interview is about the child's bedtime routine and activities the following day.

Observed in both, heat map figure 4.5 and heat map in figure 4.7 abusive transcripts starting with joy talk are often displayed as an important feature in the heat maps. Indicating that several abusive transcripts start off innocently.

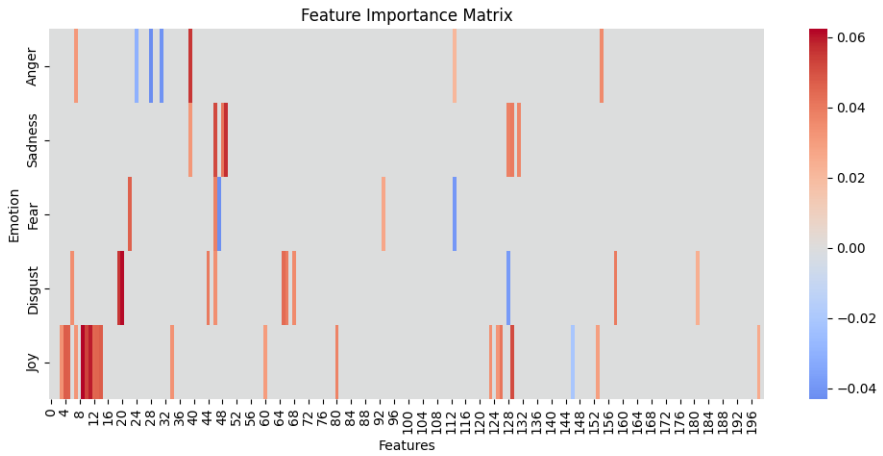

Figure 4.7: Heat map for BART prediction transcript 4 from the abusive dataset.

Feature	Sentiment	Sentence
19	Disgust	Just where he shouldn't
20	Disgust	On my private

Table 4.3: Table with sentences from transcript 4 from the abusive dataset.

BART transcript 10

Transcript 10 in the abusive dataset is only 31 utterances long, which could limit the validity of the results. The most important features are located outside this length, as seen in the heat map in figure 4.9. We see that

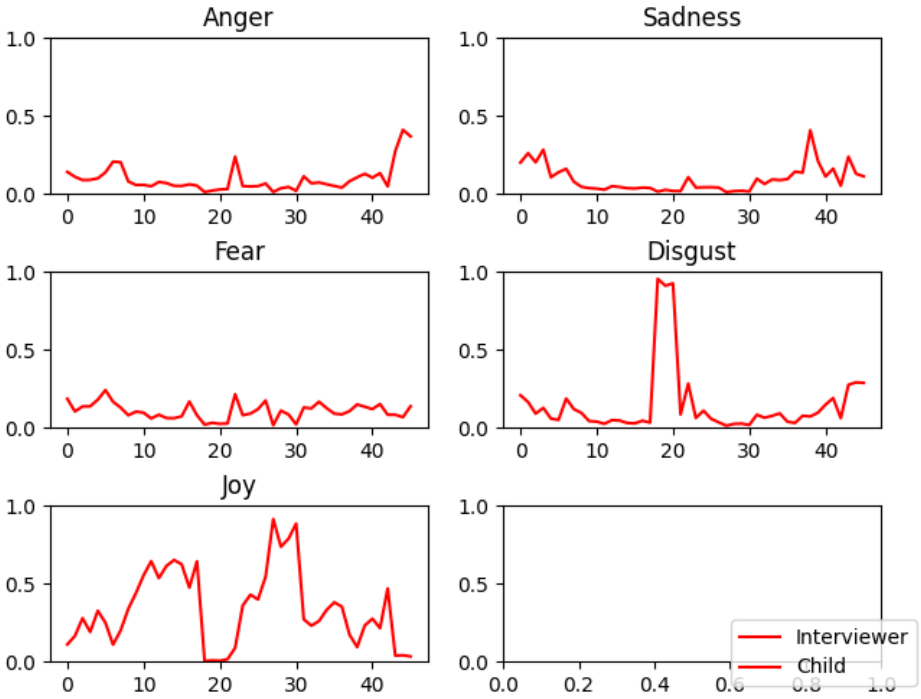

Figure 4.8: Line chart for BART prediction transcript 4 from the abusive dataset.

disgust contains several important feature groups. An especially important one is the group of important features between the ranges 127-131 for disgust. Translating this group of features into indexes in the transcript equals 2-7, they are depicted as sentences in table 4.4. Additionally, we observe two more groups of features in the row of disgust, these important feature groups also translate into the same indexes. It is interesting why the important features are most impactful towards an abusive prediction after it is repeated five times. A reason for this can be because in the heat map in figure 4.5 we observe a similar important feature group in the same location. Indicating that disgust at this index is contributing to an abusive prediction.

This transcript consists of a lot of single words as utterances. The transcript starts with the child talking to an interviewer about going to church. During the interview, the child discusses a previous instance of abuse by a priest at church. This part is where the important feature groups discussed

above are located. The abusive behavior described is slightly longer than the length of the feature groups in the heat map.

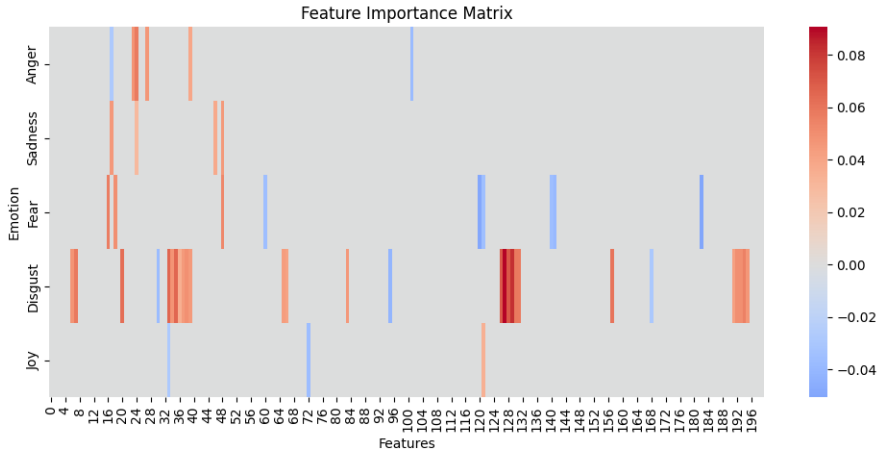

Figure 4.9: Heat map for BART prediction transcript 10 from the abusive dataset.

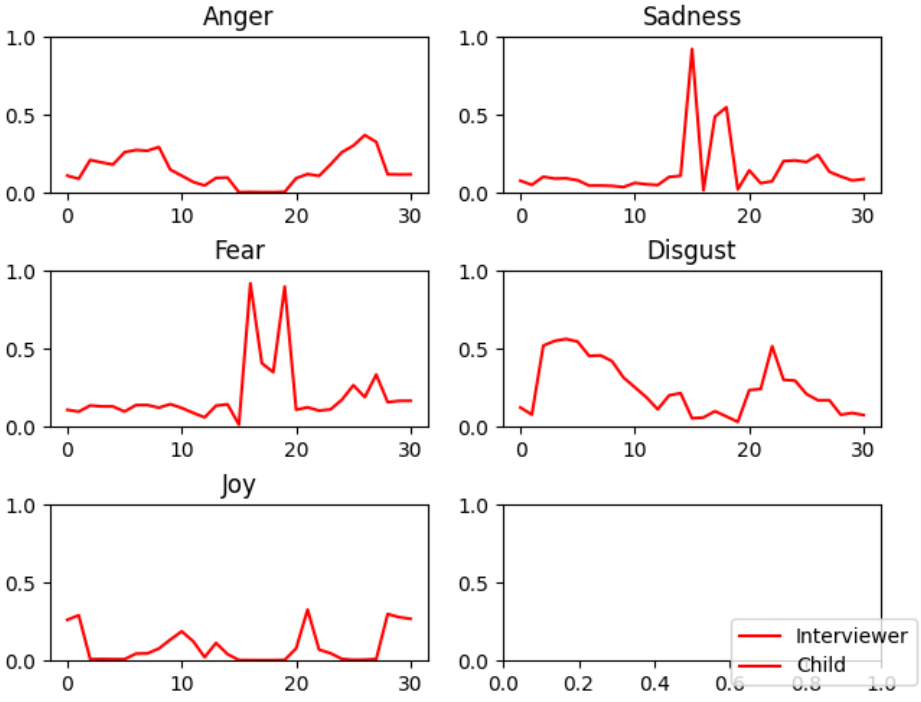

Figure 4.10: Line chart for BART prediction transcript 10 from the abusive dataset.

Index	Sentiment	Sentence
2	Disgust	Yeah I don't like the priest the man in the clothes
3	Disgust	Yeah
4	Disgust	Yeah
5	Disgust	Yeah he just did a naughty thing
6	Disgust	Yeah
7	Disgust	Yeah

Table 4.4: Table with sentences from transcript 10 from the abusive dataset.

BART transcript 12

In the heat map in figure 4.11 we can again compare the line chart in figure 4.12 local maxima with the corresponding important feature groups. Anger and fear are the most important features of the prediction of abuse. The feature groups are located between the ranges of 39-47 for fear and between the ranges of 23-30 for anger. The important feature group of anger shows similarities to the important feature group for anger in the heat map in figure 4.21.

The anger feature is particularly noteworthy in this transcript. Table 4.5 displays the sentences associated with this feature group, which describes violent and abusive behavior towards a child. On the other hand, the fear feature group appears in the transcript after the abuse has been described. A summary of the indexes in the transcript for the feature group for fear is that the child hid and was taken to the hospital.

Summary of the BART heat maps

It appears that the different sentiments contribute differently to an abusive prediction. Anger tends to be an important feature group when abuse is being described, as shown in the heat maps in figure 4.11 and the heat map in figure 4.5. On the other hand, the emotion of fear is often after the

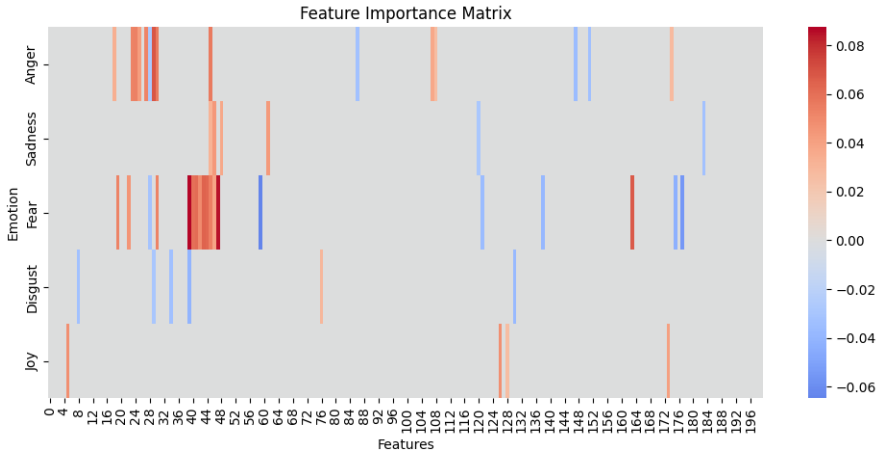

Figure 4.11: Heat map for BART prediction transcript 12 from the abusive dataset.

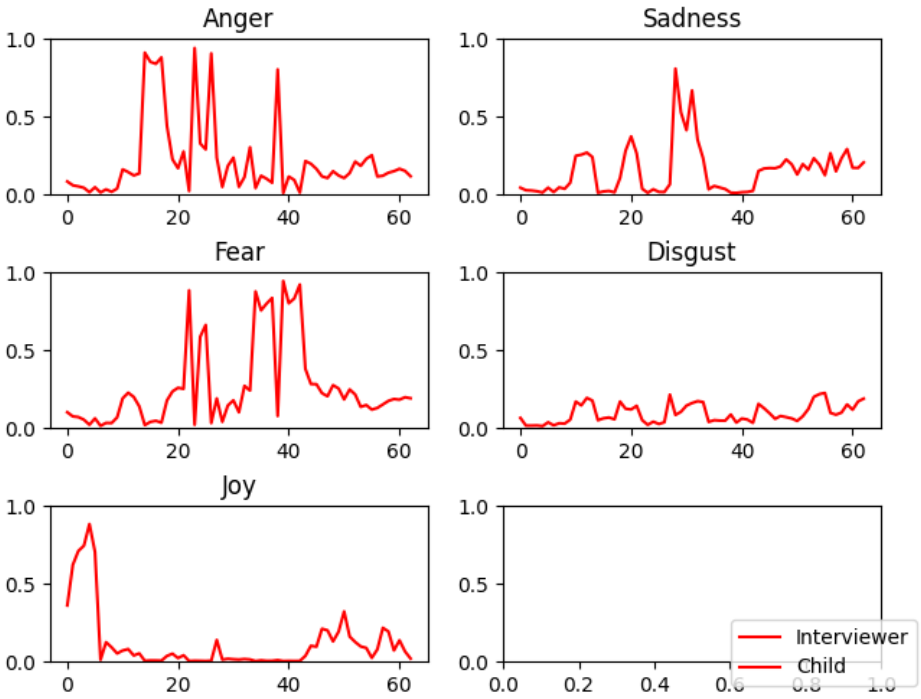

Figure 4.12: Line chart for BART prediction transcript 12 from the abusive dataset.

abuse has been described or discussed in a transcript, as seen in the heat map in figure 4.11 and in figure 4.5. Joy appears to be different from the other emotions. When joy is an important feature it is often early in the transcript, as shown in heat maps in figure 4.5 and in figure 4.7. Lastly,

Feature	Sentiment	Sentence
23	Anger	And that's when he got angry with me because I should be asleep
24	Anger	He just grabbed me and threw me
25	Anger	He just grabbed me on my chest
26	Anger	And lifted me up in the air
27	Anger	He just threw me like we were wrestling
28	Anger	But this time hurts lots
29	Anger	My mum wasn't waking up
30	Anger	She was sleeping on the floor of the kitchen

Table 4.5: Table with sentences from transcript 12 from the abusive dataset.

disgust is displayed similarly to anger and tends to be grouped during a description of abusive behavior in the transcripts.

Interestingly, the heat maps discussed above show little focus on non-abusive features, which appear randomly without any detectable pattern. However, it is worth noting that we can see some non-abusive features replacing where we have previously seen abusive features. For instance, in the heat map in figure 4.7, the features between the ranges 24-30 for anger show a small group of non-abusive features. Contrary, the heat maps in figure 4.5, and in figure 4.11 display an important abusive feature group between the same ranges.

DistilRoBERTa - transcript 12

In the heat map in figure 4.13, a group of important features is located between the ranges of 37-48 for the emotion of fear. Similarly, the line chart in figure 4.14 shows a high emotion prediction for fear within the same range. A reason for predicting fear to be so high is the sentence, "I got really scared." Because the language model DistilRoBERTa predicts by

appending previous sentences into a story, this sentence will significantly affect the whole story. These sentences are represented in table 4.6. Comparing the heat map in figure 4.13 to the heat map generated by the BART model inputs in figure 4.11, we can observe similarities between the important features for the emotion of fear between the ranges of 37-48. Demonstrating the importance of high emotion values for the indexes 40-50 in this transcript. However, the other important features group in the heat map in figure 4.11 for BART's sentiment profile, is not the same as in the heat map in figure 4.13 for DistilRoBERTa. DistilRoBERTa's heat map in figure 4.13 highlights disgust as an important feature group, while BART's heat map in figure 4.11 emphasizes the emotion of anger. This is because the BART model did not predict disgust within the ranges where the DistilRoBERTa model did as we can see from the line charts in figure 4.12 and in figure 4.14. Although both models predicted anger between the ranges of 20-28, only BART's heat map highlights this as an important feature group. It is worth noting that when disgust is predicted, it often plays a significant role in the CNN's prediction.

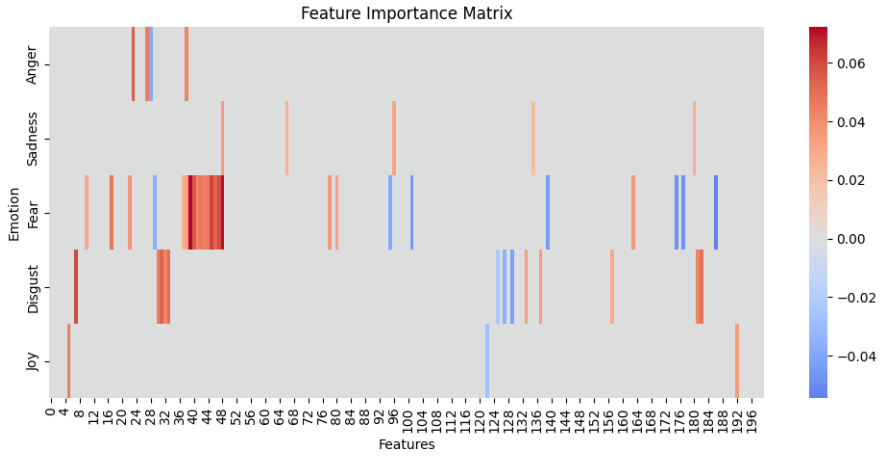

Figure 4.13: Heat map for DistilRoBERTa prediction transcript 12 from the abusive dataset.

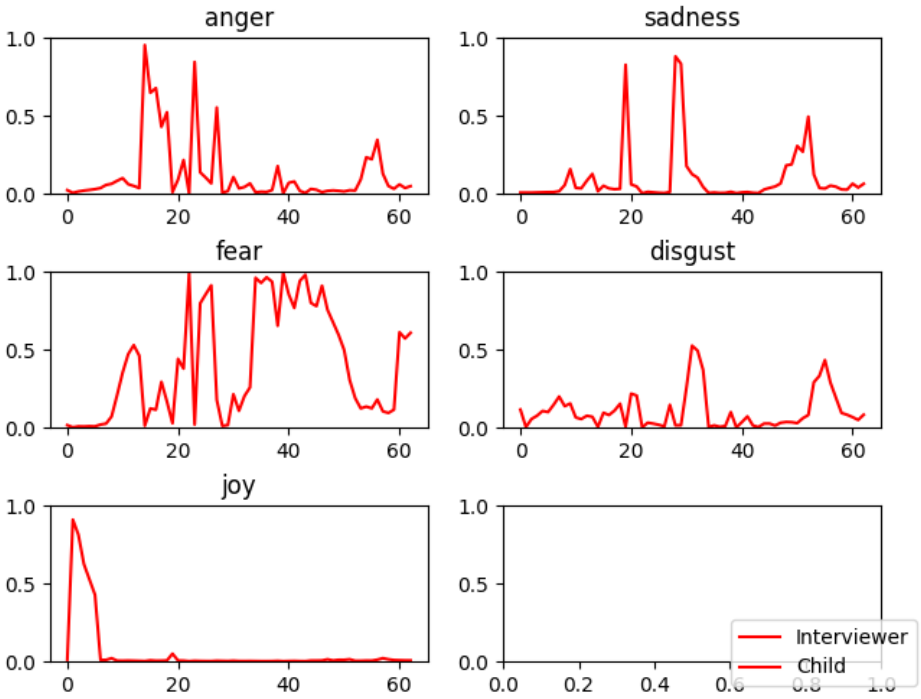

Figure 4.14: Line chart for DistilRoBERTa prediction transcript 12 from the abusive dataset.

Feature	Sentiment	Sentence
40	Fear	I went back into my room
41	Fear	And wait until he left
42	Fear	I went into my neighbour's
43	Fear	And they called the ambulance
44	Fear	They help me with my arm
45	Fear	And they took me to the hospital
46	Fear	The ambulance took Mum
47	Fear	And then my aunty picked me up from the hospital
48	Fear	The doctor told me that my arm was broken

Table 4.6: Table with sentences from transcript 12 from the abusive dataset.

DistilRoBERTa - transcript 13

In both the heat map in figure 4.15 and the line chart in figure 4.16 we see that disgust is the dominant emotion. Within this heat map, we see several groups for abusive features for the emotion of disgust. These are the features between the ranges, 6-9 and 29-32. Sentences for features 6-9 and 29-32 are represented in table 4.7. Due to the transcript's short length of only 33, padding affected the results. Therefore, features 726 - 730 represent the same index in the transcript as features 29-32. We see these indexes repeat as important features throughout the disgust row in the heat map in figure 4.15

In this transcript the child used brief phrases such as "Yeah", "I don't know", and "no". Inspecting the sentences from the features in table 4.7 reveals that the majority of the groups discussed are present in sentences that consist of more than one word.

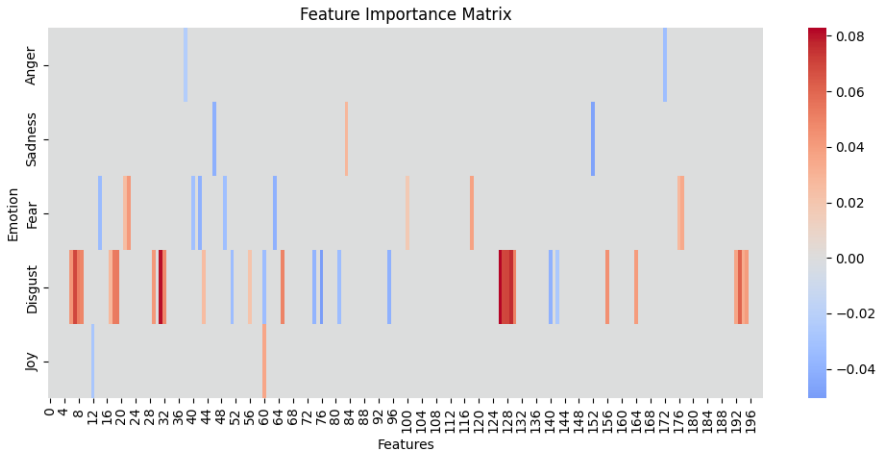

Figure 4.15: Heat map for DistilRoBERTa prediction transcript 13 from the abusive dataset.

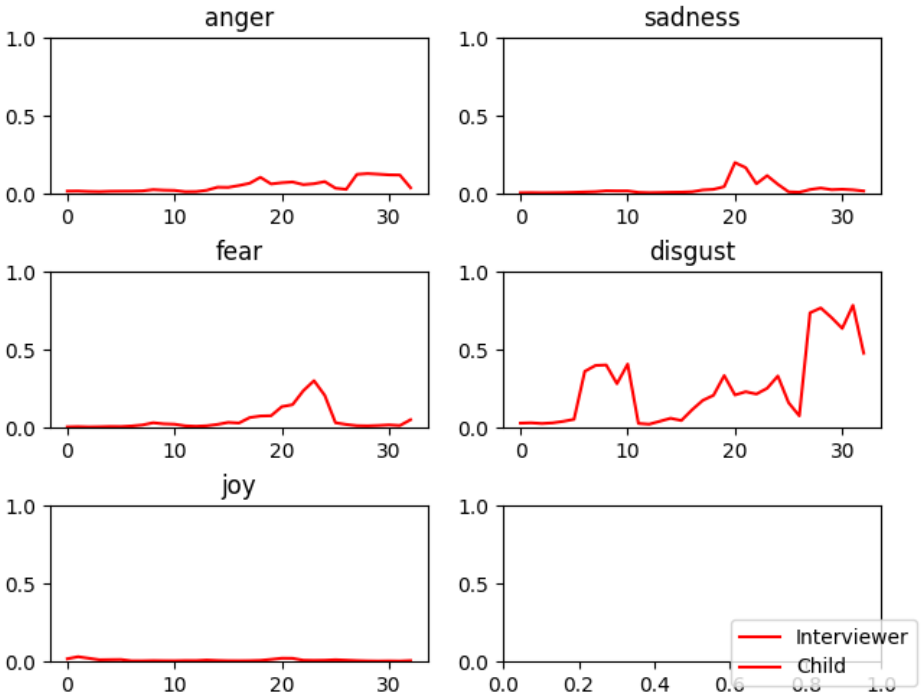

Figure 4.16: Line chart for DistilRoBERTa prediction transcript 13 from the abusive dataset.

Feature	Sentiment	Sentence
6	Disgust	Just on my privates
7	Disgust	Yeah
8	Disgust	Just a lot of times
9	Disgust	I don't know
29	None	The last time there was this boy and I have to teach him how to play the big boys' game
31	Disgust	Sam told me that I have to teach him how to play
32	Disgust	And he made me touch his pee-pee

Table 4.7: Table with sentences from transcript 13 from the abusive dataset.

DistilRoBERTa - transcript 20

Transcript 20 In the abusive dataset is quite small, containing only 35 utterances. This causes the features to be represented in the padding of the sentiment profile. Upon examining the heat map in figure 4.17, we can identify three groups of features that contribute to an abusive prediction: disgust features between the ranges of 7-11 and 16-19, and fear features between the ranges of 39-41. The heat map displays similar feature groups as previous heat maps examined, namely in table 4.13 and 4.15. This suggests that certain sections within a transcript contribute towards an abusive prediction. The features in the heat map in figure 4.17, within these sections, are displayed in table 4.8. These sentences clearly exhibit abusive behavior. Upon examining the transcript, it showed that almost all of the abusive sentences in this transcript are featured in the heat map in figure 4.17.

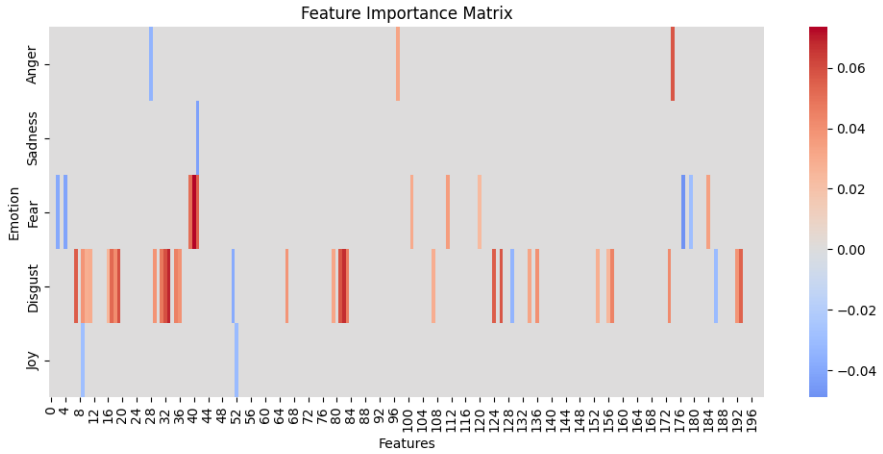

Figure 4.17: Heat map for DistilRoBERTa prediction transcript 20 from the abusive dataset.

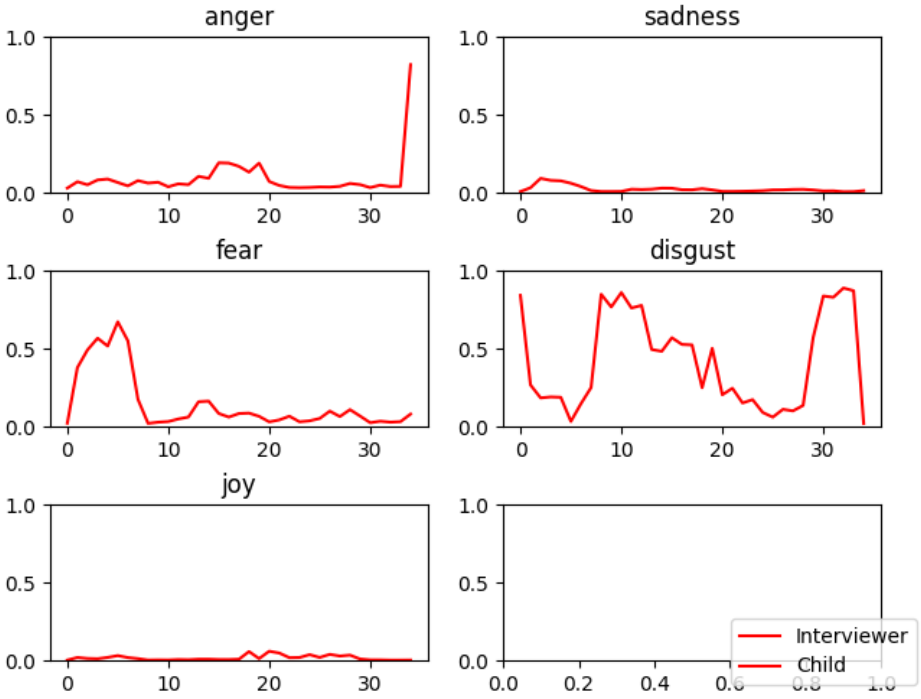

Figure 4.18: Line chart for DistilRoBERTa prediction transcript 20 from the abusive dataset.

Feature	Sentiment	Sentence
7	Disgust	When I went into the cupboard he followed me in and told me he would teach me a special show and tell game and that's when he pulled his pants down
8	Disgust	Then he stood there and his willy was big and yuck
9	Disgust	Then he told me it was my turn to go
10	Disgust	Then I had to pull my dress up and my undies down
11	Disgust	I started crying and he gave me a hug
16	Disgust	He put both his arms around me and then he told me that I could go outside
17	Disgust	Well then I went outside and played on the playground with my friends
18	Disgust	He said that it was a special show and tell game and he was teaching me because I was a good helper
19	Disgust	And that's when he pulled his pants down

Table 4.8: Table with sentences from transcript 20 from the abusive dataset.

Summary of the DistilRoBERTa abusive heat maps

After analyzing several heat maps, including the heat map in figure 4.13, the heat map in figure 4.15, and the heat map in figure 4.17, it's evident that they all showcase similar patterns of important features. It appears that the DistilRoBERTa heat map's important features are mostly the emotions of fear and disgust. Moreover, the features are often located between the ranges 4-10 and 28-36 for the emotion of disgust and between the ranges 36-48 for the emotion of fear. It seems that non-abusive features are few and far between, possibly due to the short transcripts which lack extended periods of non-abusive conversation. Based on the transcripts selected for DistilRoBERTa's heat maps, it appears that there were numerous instances of abusive discussions, with short sections of non-abusive conversation. Consequently, it's anticipated that there will be limited non-abusive attributes.

4.2.2 Results from non-abusive transcripts

In this section, we will discuss the outcomes of the non-abusive heat maps. When the importance values are negative, it implies that the feature is non-abusive. The heat maps were generated using DistilRoBERTa's and BART's predicted sentiment profiles. The non-abusive features are displayed as blue in the heat maps. Our attention will be on these features, and examine if there are any specific patterns or reasoning behind the CNNs decision of non-abusive prediction.

DistilRoBERTa

In the heat map in figure 4.19 there are two noticeable patterns of blue feature values, these are between the ranges 31-36, and 138-141. The sentences from features 31-36 are listed in table 4.9. Upon inspection, these sentences does not tell us much, except it does not contain any abusive behavior. As shown in previous heat maps, sections containing disgust between ranges 31-36 often result in an abusive prediction. This could

explain why these ranges exhibit a blue group of features.

Similarly, upon examining the heat map in figure 4.20, no noteworthy observations were made. Analyzing the sentences associated with the top 8 features listed in table 4.10 no clear reasoning was found as to why the prediction was non-abusive. The sentences and words contain random words and phrases that contribute to different predictions. The main conclusion that can be drawn from both the non-abusive heat maps discussed is that there is an insignificant amount of red feature groups present.

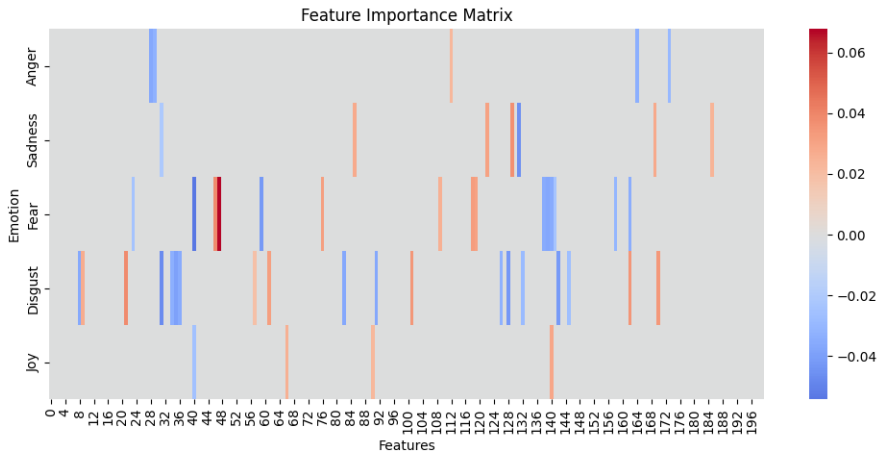

Figure 4.19: Heat map for DistilRoBERTa prediction transcript 2 from the non-abusive dataset.

BART

Upon analyzing the heat maps generated by the sentiment profiles from the BART language model, we observed similar findings to those from DistilRoBERTa. The main takeaway from the heat map in figure 4.21 and heat map in figure 4.22 is the absence of abusive features we identified in the abusive heat maps, such as fear (features between 40-50) and disgust (features between 25-35). Apart from that, the non-abusive features seem to be scattered randomly across the entire heat map. We will discuss this in more detail in the next chapter.

Feature	Sentiment	Sentence
31	Disgust	They were in my stocking. I think mummie must've gave them to me
32	Fear, Joy	And um, lego.
33	Fear	And i got and um some track with a car that goes round lots of loops
34	Fear	Yes.
35	Fear	Sometimes goes like that.
36	Fear	Or it can go like that.

Table 4.9: Table for DistilRoBERTa prediction transcript 2 from the non-abusive dataset.

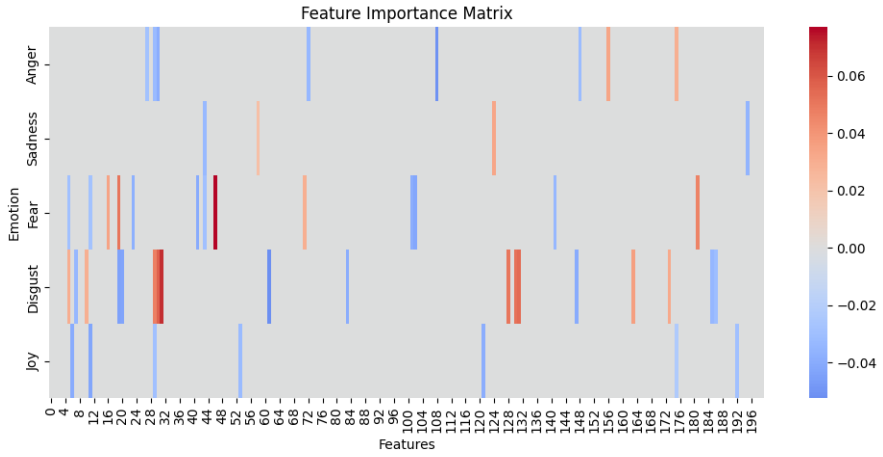

Figure 4.20: Heat map for DistilRoBERTa prediction transcript 15 from the non-abusive dataset.

4.3 Comparison of two abusive heat maps

The heat maps in figure 4.23 and the heat map in figure 4.24 for transcript 7 in the abusive dataset display similarities in groups of important features. Specifically, features 30-38 indicate disgust, features 40 and 41 indicate fear, and features 22-24 indicate anger. The sentences for the disgust features can be found in table 4.11. The transcript revolves around a child

Feature	Importance value	Sentence
446	0.077	The dog
631	0.070	Can i go to the toilet
630	0.057	Lamp
731	0.054	Postman
661	-0.052	And um is one window is putting
730	0.052	He's going to a match
419	0.052	Can like that
108	-0.051	That goes in the kitchen

Table 4.10: Table for top 8 features in transcript 15 in the non-abusive database (Red is an abusive feature, blue is a non-abusive feature)

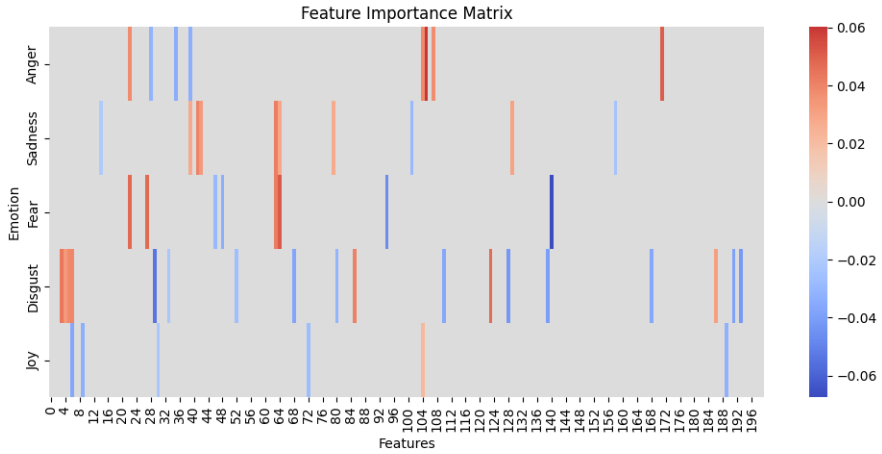

Figure 4.21: Heat map for BART prediction transcript 1 from the non-abusive dataset.

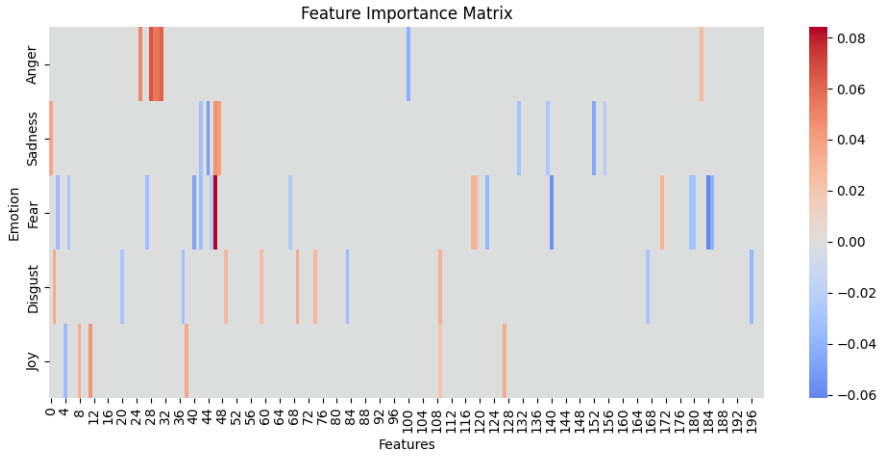

Figure 4.22: Heat map for BART prediction transcript 2 from the non-abusive dataset.

playing Jenga with their cousin, who suddenly starts hitting them. This is where the important feature groups are located. Almost every abusive important feature group from one heat map can be found in another. This could suggest that even though different language models predict different confidence scores for emotions. They are still high enough to be considered important. Therefore, by combining the confidence score of language models, we could achieve better sentiment analysis based on these patterns of important features.

Feature	Sentiment	Sentence
30	Disgust	Five hundred
32	Disgust	That he was mean to me
33	Disgust	He stopped when I was bleeding
34	Disgust	And I just went to the toilet
35	Disgust	No
37	Disgust	Yeah
40	Fear	And he hit me
41	Fear	He's my cousin

Table 4.11: Table with sentences from transcript 7 from the abusive dataset.

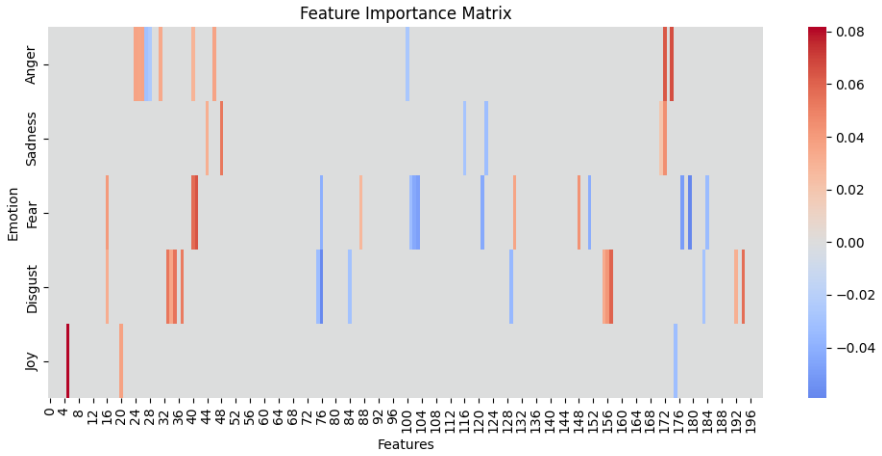

Figure 4.23: Heat map for BART prediction transcript 7 from the abusive dataset.

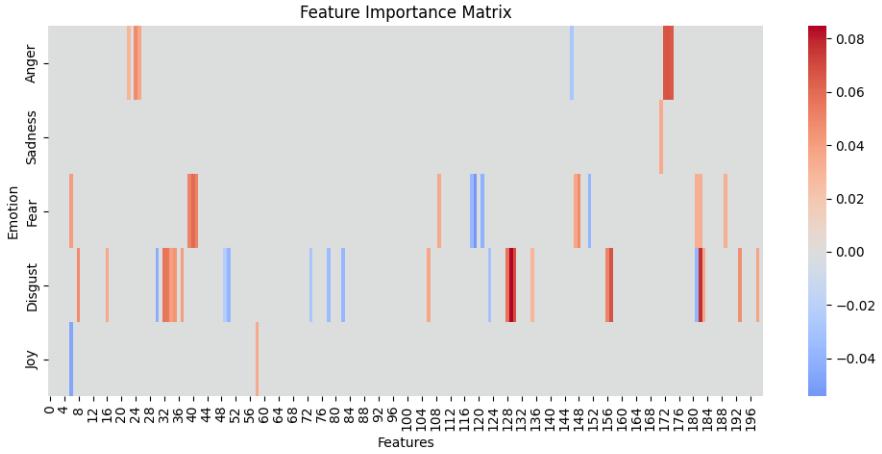

Figure 4.24: Heat map for DistilRoBERTa prediction transcript 7 from the abusive dataset.

In subsection 4.2.1, we examined the BART model's heat map for transcript 10 in the abusive dataset. We observed that index 2 to index 7 in the transcript were repeatedly highlighted as important features for the prediction of the CNN. Similar patterns are observed in the heat map generated by DistillRoBERTa's heat map in figure 4.26. Additionally, comparing the row of anger and fear between the two heat maps shows similarities. This indicates that the CNN's prediction decision is influenced by the input from both models. When we benchmarked the two models,

we found that they differed in the emotion they predicted and the score assigned to the other emotions. DitstilRoBERTa predicted significantly more disgust than BART. On the contrary, BART predicted more joy than DistilRoBERTa. By analyzing the heat maps generated by both models, we can discover similarities in their prediction of sentiment scores. We will explore this further in the next chapter.

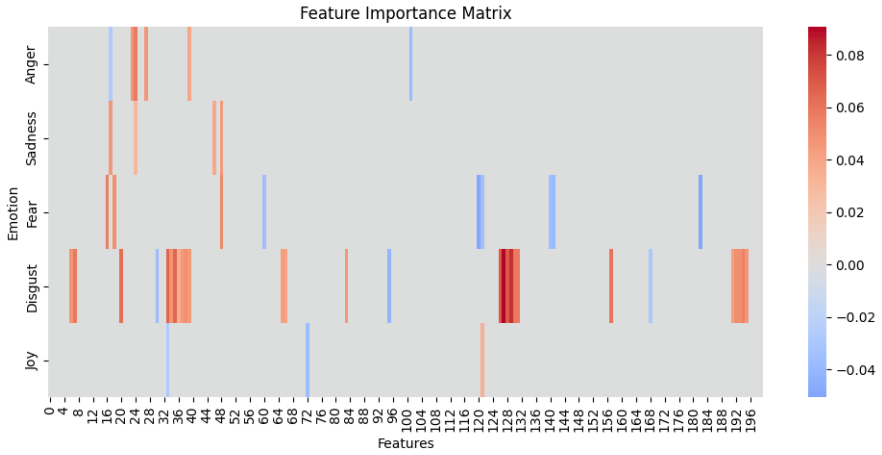

Figure 4.25: Heat map for BART prediction transcript 10 from the abusive dataset.

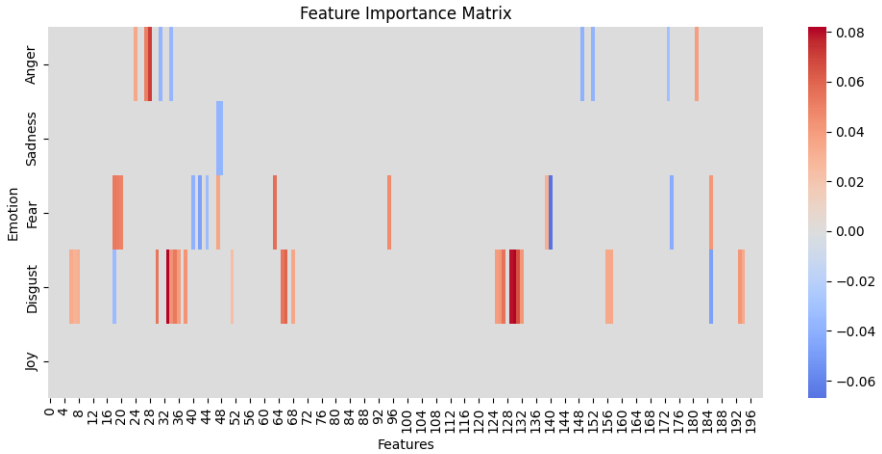

Figure 4.26: Heat map for DistilRoBERTa prediction transcript 10 from the abusive dataset.

4.4 Summary

In this chapter, we analyzed the heat maps and compared them to each other. Section 4.1 explained the metrics and the heat maps' design. Moreover, explained what importance of values and important features are. The section concluded with an explanation of why we chose to eliminate the neutral sentiment from the heat maps.

In section 4.2 we analyzed the heat maps generated from the abusive and the non-abusive datasets. The models used were BART and DistilRoBERTa, and the heat maps were based on the emotion analysis of the Fletcher and Mock interviews datasets, which consisted of 80 transcripts together. The sentiment profile from both models was used for both training and feature extraction. LIME [8] was used for feature extraction, and a CNN was employed as the DNN due to the sentiment profiles being inserted as images. The results showed that certain patterns are indicative of abusive prediction, such as anger features 20-32, fear features 36-48, and disgust features 4-10 and 28-36. On the contrary, non-abusive heat maps did not contain any repeating patterns, leading us to conclude that the CNN's prediction of non-abusive data is based more on the absence of abusive patterns.

In section 4.3 we compared heat maps from same transcripts in the abusive dataset. Despite differences in the sentiment profiles of the models used, the heat maps in this section displayed similarities between important features for the prediction of abusive.

The next chapter will analyze the meaning behind these results. We will also discover additional experiments for improving our understanding of the prediction made by a DNN.

Chapter 5

Discussion

In the previous chapter, the abusive and non-abusive heat maps were explored and studied. We saw that certain regions in the heat maps for the emotions were important for the CNNs prediction. This chapter will provide a more comprehensive overview of the findings. Firstly, we will discuss the heat maps generated from the two language models' sentiment profiles. Secondly, we will examine different uses for the extracted features. Thirdly, we will discuss the validity of the results achieved and address some difficulties encountered. Lastly, we will introduce potential solutions.

5.1 Importance of sentiments

To better understand the decision-making process of DNNs, we used a CNN to analyze the sentiment profiles provided by the two language models. This helped determine which emotions were the most significant in identifying abusive transcripts. For example, the heat map in figure 4.25 and the heat map in figure 4.26 displayed similar clusters of important abusive features for the emotions; fear, anger, and disgust. Indicating that for transcript 10 in the abusive dataset, these three emotions contributed heavily toward the prediction of abusive. When translating the group of abusive features to indexes in the transcript we were often able to locate

where in the transcript the abusive actions were described. In contrast, the non-abusive features rarely contributed information as to why it was predicted as non-abusive. The main noticeable pattern of non-abusive features was either lack of abusive feature groups or a majority of non-abusive features spread across the heat map. A reasoning behind this is that when labeling something as "not", it is often difficult to extract certain features as to why. To illustrate this, we can inspect an example of trying to differentiate a sunflower from other flowers. The main features for the class of "other flowers" would be the absence of features typical in a sunflower. This is similar to the results we observed in the previous chapter for non-abusive transcripts.

Interestingly, we saw shared importance for emotions from both models when predicting abusive transcript. When one language model labeled a sentence with a lower confidence score than the other for a certain emotion. It still occurred that the sentence was important for a prediction. Inspecting the line chart in figure 5.5 for transcript 10 in the abusive dataset we can see that fear and sadness are the emotions with the highest confidence scores. However, for the generated heat map in figure 4.9 using this sentiment profile as input, displays disgust as the most important feature for predicting abuse. Even though disgust has a lower confidence score in the line chart in figure 5.5. This shows that combining different sentiment profiles shifts the important features to where they have similar confidence values. An experiment to strengthen this hypothesis is generating heat maps with only one model's sentiment profile as input. This resulted in losing the clusters of important features observed earlier and the importance value of each feature decreased. An example of this is when only using the BART model's sentiment profile as input to generate the heat map as in the heat map in figure 5.1. In the heat map 5.1 the groups of abusive features are less noticeable. Moreover, most of the important features appear much more spread. When inspecting the heat map for the same transcript in figure 4.7, we see much more preciseness in where

the prediction is decided. Both the heat maps used the same sentiment profile displayed in the line chart in figure 4.8. We can see how the heat maps match the sentiment profile much better for the heat map in figure 4.7 where both models' sentiment profile was used.

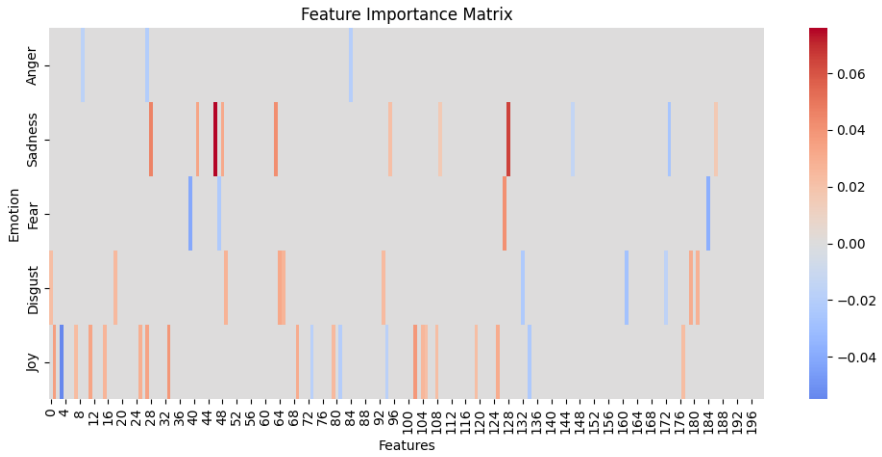

Figure 5.1: Heat map from transcript 4 in the abusive dataset generated using only BART's sentiment profiles

5.2 Using the important features

As discussed in the previous section, heat maps from the same transcripts displayed shared important features, even though the predicted confidence score was lower for one. By using these important features to inspect the difference in confidence values, we could cascade language models for more accurate sentiment analysis. For example the heat map in figure 4.23 and the heat map figure 4.24 we see that disgust has a group of important features. When examining the line chart in figure 5.3 we can see that the emotion of disgust does not have a high confidence score for the group of important features. If we compare it to the line chart in figure 5.2 we see a much higher confidence score for disgust. Because of this, we can draw the conclusion that the emotion of disgust is more important than initially displayed in BART's line chart for abusive transcript 7 in figure 5.3. By inspecting the range of important features in heat maps, we can use

it to cascade language models to understand children's emotions better. Using the range of important features to cascade a language model using an abusive transcript and increase the confidence score of disgust for these predictions. We see similarities for transcript 10 in the abusive dataset as well. The emotion of disgust has a higher confidence score in the line graph in figure 5.2 than in the line chart in figure 5.3 within the important feature group for disgust in the heat map in figure 4.25 and in the heat map in figure 4.26.

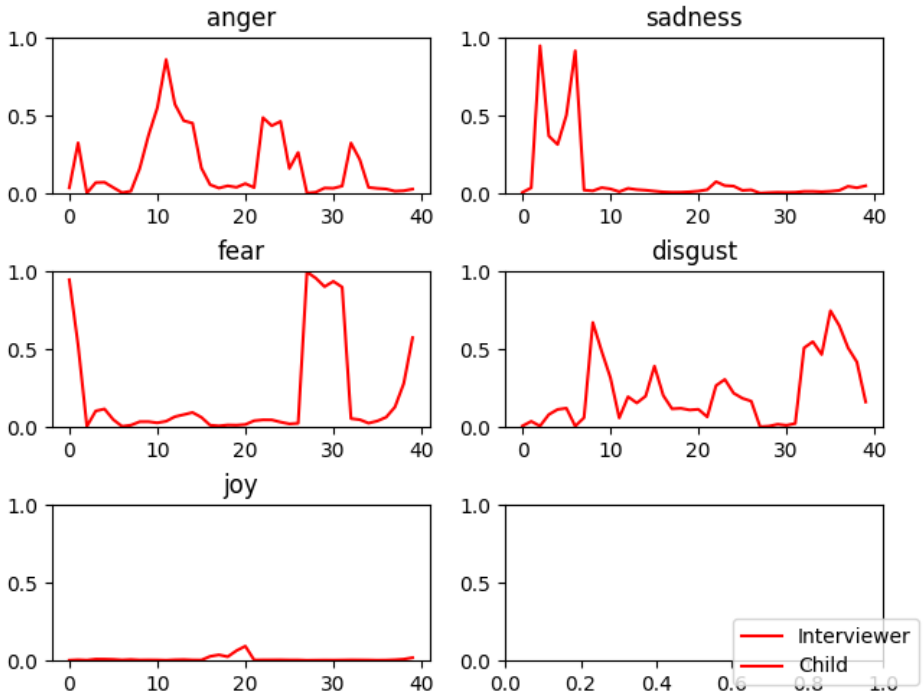

Figure 5.2: Line chart for DistilRoBERTa prediction transcript 7 from the abusive dataset.

Another use of important feature groups is to design a way to conduct abusive investigative interviews. Using abusive transcripts from interviews with a high succession rate for gaining information as input and examining the patterns of emotions. Inspect if a certain way of conducting investigative interviewing achieves more information. Create a basic template as to how to conduct investigative interviewing. With different transcripts, it could possibly be translated to different fields as well. An ethical issue with this though, is that we are manipulating emotions to gain more productive

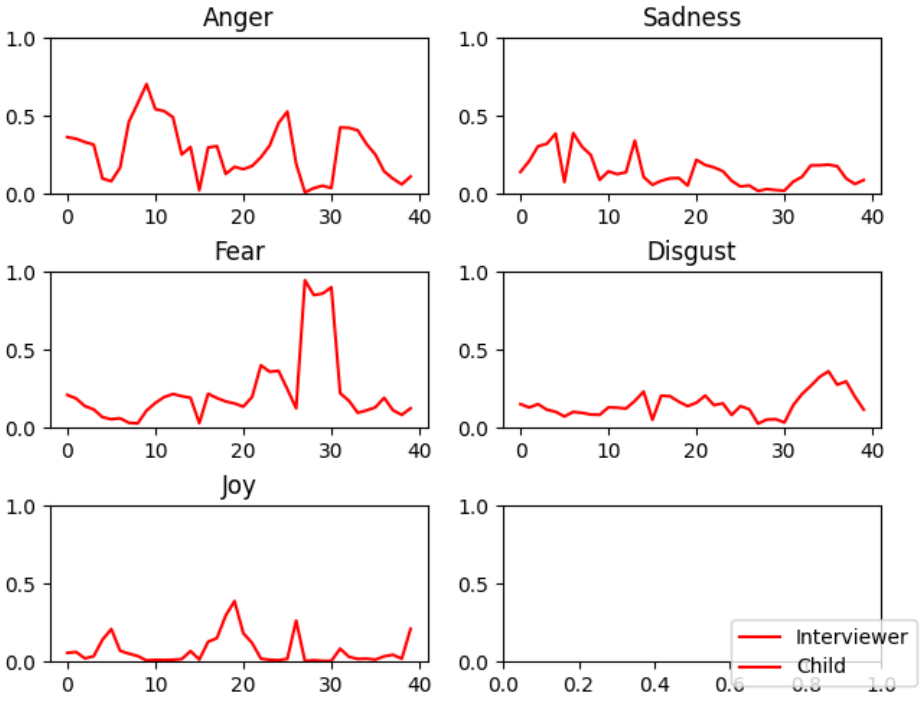

Figure 5.3: Line chart for BART prediction transcript 7 from the abusive dataset.

interviews.

5.3 Validity of the results achieved

When the CNN was tested on the test set it achieved an accuracy of 100% corerect predictions. However, this may not necessarily be a good sign. It could be that the difference in confidence scores is too large, and the CNN is merely detecting this distinction. Alternatively, the CNN prediction is based on recurring patterns. This is because every input from the sentiment profile is repeated until it reaches a length of 1000 confidence scores. As the thesis revolves around, understanding the decisions made, it is not helpful to learn that the CNNs decision is based on repeating a pattern. The main pattern observed in the non-abusive heat maps was the absence of abusive feature groups. It was difficult to extract any repeating patterns of non-abusive features. This makes it difficult to understand the reason behind the non-abusive prediction. However, it is worth noting that

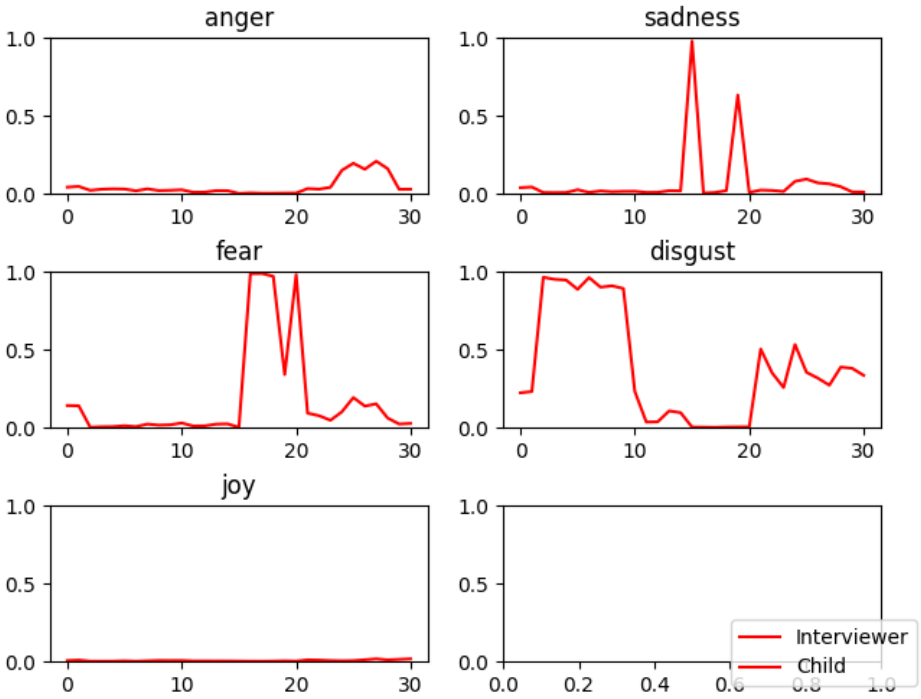

Figure 5.4: Line chart for DistilRoBERTa prediction transcript 10 from the abusive dataset.

the non-abusive transcripts contain conversations about different topics. So it would be difficult for the CNN to put emphasis on a non-abusive feature as touched upon earlier in this chapter. Therefore, the focus of this work was mostly on the abusive features. The feature extraction for abusive transcripts did return patterns and sentences. Hence, increasing our understanding of as to how DNNs work.

5.4 Challenges

One of the challenges encountered during the experiments was the limited availability of abusive transcripts. For non-abusive transcripts, there were a lot of available transcripts to gather into a dataset. Additionally, the length of the abusive transcripts was around an average of 50 utterances compared to the average of 200 utterances for the non-abusive transcripts. For the CNN to predict, a padding had to be added to the end of each transcript in the abusive dataset. The max length for each transcript was

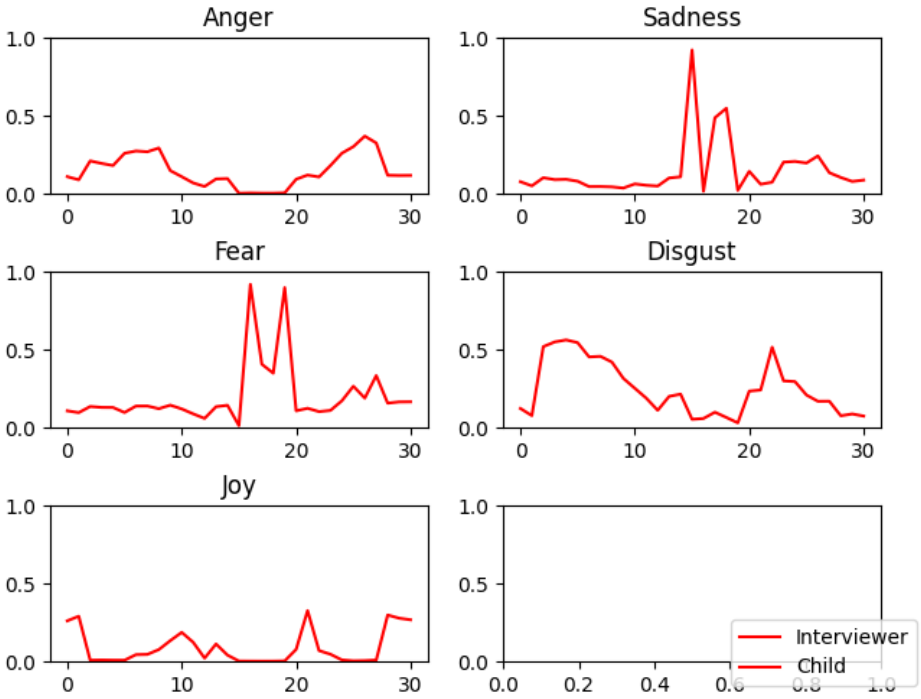

Figure 5.5: Line chart for BART prediction transcript 10 from the abusive dataset.

200. A few of the non-abusive transcripts had to be shortened because they were too long. An additional challenge encountered is that most of the language models available for sentiment analysis were either difficult to use or not applicable to this thesis. The thesis scope was extracting the emotional score from children. We were not able to find any language models fine-tuned for classifying children's emotions. Additionally, we did not have an available dataset for which we could fine-tune a language model ourselves.

5.5 Future experiments for improved results

Additional experiments can be conducted to improve the understanding of how DNNs decide their outcome and to see if a transcript can be separated purely using emotional scores. The experiments proposed below are a solution to improve both the validity of the results and the challenges encountered

1. Larger number of improved transcripts.

2. Include extra language models.

3. Multi-class predictions.

4. Use a dictionary-based approach instead of sentiment profiles.

Improving the datasets by increasing the number of transcripts and gathering longer transcripts. As explained earlier, all the abusive transcripts had to be padded to 200 in length. This resulted in earlier sections of the transcripts mostly being featured by LIME [8]. Moreover, for both the abusive and non-abusive datasets, the input was only 40 transcripts. Nowadays, there are several models with large datasets. The model could learn and train better with an increased number of transcripts for the datasets.

There were only two models included as input into the CNN. Instead of increasing the number of transcripts, the number of language models could be increased. The results in section 3 show that the language models produced different scores for different sentiments. Several existing datasets have been labeled through humans voting for the correct sentiment. Using several language models to train the CNN, the correct sentiment is chosen as a majority vote.

Using multi-class predictions we can also extract features typical for other types of interviews. This could be beneficial to separate the abusive features even more. However, it could also result in the opposite. This is because different types of interviews might share similar traits as abusive ones.

The last proposed experiment is to avoid using a sentiment profile as features. Implement a dictionary containing all the words from the sentences in the transcripts. We predict sentiment for each sentiment and then split the sentence into words. Each word in that sentence is assigned a sentiment. Doing this for every sentence we gain a unique word-to-

sentiment profile which we can insert into the CNN for analysis. The result of this could show common words for abusive transcripts. However, this would increase the size of features of the heatmap to a significantly higher number. To avoid this, some words could be eliminated, for example, common words like “is”, “a”, “this”, “the”, etc. Additionally, we would need a padding because each transcript does not contain the same amount of words. This padding word has to be ignored by the CNN. Hopefully, the heat map generated with this approach could yield feature words usually uttered in abusive contexts.

Chapter 6

Conclusion

6.1 Summary

In this thesis, we used explainable AI to understand the inner workings of a DNN, within the context of children's emotions in the interviews. We short-listed different available language models to predict sentiments in a conversational data. As we used the language model predictions to create sentiment profiles of the child throughout a transcript, we benchmarked their ability to predict children's emotions. The benchmarking was applied to an emotion-annotated dataset containing conversations between a child and a professional interviewer. Results showed that BART and DistilRoBERTa language models perform best at creating sentiment profiles. Before creating sentiment profiles we had to obtain data representing both abusive and non-abusive. We obtained a non-abusive dataset by scraping interviews between a child and a trained professional in the Fletcher project [59]. An abusive dataset was obtained from the Centre for Investigative Interviewing[1]. The sample size of the abusive and the non-abusive datasets was only 20 transcripts each. With the two datasets, the language models predicted confidence scores for each of six emotions for every sentence in the transcripts. This created an image

[1]https://www.investigativecentre.com/

of the child's emotional state throughout a transcript which we inserted into a CNN. We explained the prediction of the CNN when classifying a sentiment profile using LIME. We then plotted the important features onto a heat map for further analysis and examination.

6.2 Main Contributions

In this section, we revisit the research questions in the problem statement outlined in section 1.2. The aims are listed below and we look at achievements and information gathered for solving them.

1. Can explainable AI be used to uncover the decisions in deep learning models? In this thesis, we used LIME to uncover the decision of the DNN [8]. Utilizing LIME, we extracted the important features of the CNNs decision. When inserting these important features into a heat map, we observed that the features extracted corresponded to the abusive sections within a transcript. However, when inspecting the non-abusive features, we gained little to no understanding as to why. This caused us to further analyze by inserting only sentiment profiles from a single language model, revealing lesser noticeable abusive features and a lower importance value. Hence, displaying that combining the language model's sentiment profiles improved the accuracy of where abusive actions are transcribed. By increasing our trust in CNNs we can apply the information gained to future experiments to enhance existing technologies.

2. Are there certain patterns of dialogue within abusive and non-abusive interviews?

As mentioned in the previous answer, we observed patterns repeating across different heat maps. Moreover, these patterns often covered where the abusive sections were in a transcript. Using these results we could achieve better sentiment predictions by language models, through cascading [72]. Inspect the areas where the important features are located and use the language models' difference in confidence values to improve

sentiment classification.

3. Can emotion analysis be used to classify the type of transcript

As argued in chapter 5, answering this research question requires additional experiments. We saw that the CNN received an accuracy of 100%. However, the extracted features for the non-abusive did not reveal any certain patterns or sections. We did not enhance our understanding of DNNs.

The overall aim of the thesis was to explore the prediction of a DNN using explainable AI to analyze the prediction of a DNN with sentiment profiles generated by language models as input. We argue that we observed interesting information which could be used in several future experiments and enhancements of existing technologies. The extracted abusive features contained valuable information related to the abusive actions described. Using purely emotions as input it is interesting that we can locate these sentences in the transcript.

6.3 Future Work

As future work is discussed in more detail in section 5.5, we briefly mention the main idea behind it, larger number of improved transcripts, include more sentiment profiles predicted by language models, use multi-class predictions instead of only 2 classes, and use a dictionary-based approach instead of sentiment profiles.

Bibliography

[1] Michael E. Lamb and Deirdre A. Brown. "Conversational apprentices: Helping children become competent informants about their own experiences." In: *British Journal of Developmental Psychology* 24.1 (Mar. 2006), pp. 215–234. DOI: 10.1348/026151005X57657.

[2] Lisa Rudolfsson. ""At Least I Tried": Swedish Police Officers' Experiences of Meeting with Women Who Were Raped." In: *Journal of Police and Criminal Psychology* 37.2 (Feb. 2022), pp. 365–376. DOI: 10.1007/s11896-021-09435-0.

[3] Pegah Salehi et al. "Synthesizing a Talking Child Avatar to Train Interviewers Working with Maltreated Children." In: *Big Data and Cognitive Computing* 6.2 (June 2022), p. 62. DOI: 10.3390/bdcc6020062.

[4] Syed Zohaib Hassan et al. "A Comparative Study of Interactive Environments for Investigative Interview of A Virtual Child Avatar." In: *2022 IEEE International Symposium on Multimedia (ISM)*. 2022 IEEE International Symposium on Multimedia (ISM). Dec. 2022, pp. 194–201. DOI: 10.1109/ISM55400.2022.00043.

[5] Syed Zohaib Hassan et al. "A Virtual Reality Talking Avatar for Investigative Interviews of Maltreat Children." In: *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*. New York, NY, USA: Association for Computing Machinery, Oct. 7, 2022, pp. 201–204. DOI: 10.1145/3549555.3549572.

[6] Mairi Benson and Martine Powell. "Organisational challenges to delivering child investigative interviewer training via e-learning." In:

International Journal of Police Science & Management 17.2 (June 1, 2015), pp. 63–73. DOI: 10.1177/1461355715580912.

[7] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier." In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[9] P.J. Denning et al. "Computing as a discipline." In: *Computer* 22.2 (1989-02), pp. 63–70. DOI: 10.1109/2.19833.

[10] D. E. Comer et al. "Computing as a discipline." In: *Communications of the ACM* 32.1 (Jan. 1989). Ed. by Peter J. Denning, pp. 9–23. DOI: 10.1145/63238.63239.

[11] Ulrika Östlund et al. "Combining qualitative and quantitative research within mixed method research designs: A methodological review." In: *International Journal of Nursing Studies* 48.3 (2011), pp. 369–383. ISSN: 0020-7489. DOI: https://doi.org/10.1016/j.ijnurstu.2010.10.005.

[12] Luke Stark and Jesse Hoey. "The Ethics of Emotion in Artificial Intelligence Systems." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada, Mar. 3, 2021, pp. 782–793. DOI: 10.1145/3442188.3445939.

[13] Michael E. Lamb et al. "Structured forensic interview protocols improve the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol." In: *Child abuse & neglect* 31.11 (2007), pp. 1201–1231. ISSN: 0145-2134. DOI: 10.1016/j.chiabu.2007.03.021. URL: https:

//www.ncbi.nlm.nih.gov/pmc/articles/PMC2180422/ (visited on 05/28/2023).

[14] Gunn Astrid Baugerud et al. "Multimodal Virtual Avatars for Investigative Interviews with Children." In: *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval*. Taipei, Taiwan: Association for Computing Machinery, 2021, pp. 2–8. DOI: 10.1145/3463944.3469269.

[15] Gunn Astrid Baugerud et al. "Forensic interviews with preschool children: An analysis of extended interviews in Norway (2015-2017)." In: *Applied Cognitive Psychology* 34.3 (2020), pp. 654–663. DOI: 10.1002/acp.3647. URL: https://www.duo.uio.no/handle/10852/80204.

[16] Yael Karni-Visel et al. "Facilitating the Expression of Emotions by Alleged Victims of Child Abuse During Investigative Interviews Using the Revised NICHD Protocol." In: *Child Maltreatment* 24.3 (Aug. 2019), pp. 310–318. ISSN: 1077-5595, 1552-6119. DOI: 10.1177/1077559519831382. URL: http://journals.sagepub.com/doi/10.1177/1077559519831382 (visited on 05/29/2023).

[17] Xueming Luo et al. "Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases." In: *Marketing Science* 38.6 (Nov. 2019), pp. 937–947. DOI: 10.1287/mksc.2019.1192.

[18] Endang Wahyu Pamungkas. "Emotionally-Aware Chatbots: A Survey." In: *Computing Research Repository* abs/1906.09774 (2019).

[19] Eleni Adamopoulou and Lefteris Moussiades. "Chatbots: History, technology, and applications." In: *Machine Learning with Applications* 2 (Dec. 15, 2020), p. 100006. DOI: 10.1016/j.mlwa.2020.100006.

[20] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. "A Survey on Conversational Agents/Chatbots Classification and Design Techniques." In: *Web, Artificial Intelligence and Network Applications*. Cham, 2019, pp. 946–956. DOI: 10.1007/978-3-030-15035-8_93.

[21] Bayan Abu Shawar and Eric Atwell. "Chatbots: Are they Really Useful?" In: *Journal for Language Technology and Computational Linguistics* 22 (2007), pp. 29–49.

[22] Diksha Khurana et al. "Natural language processing: State of the art, current trends and challenges." In: *Multimedia tools and applications* 82.3 (2023), pp. 3713–3744.

[23] *NLP vs. NLU vs. NLG: the differences between three natural language processing concepts*. Nov. 12, 2020.

[24] Prissadang Suta et al. "An overview of machine learning in chatbots." In: *International Journal of Mechanical Engineering and Robotics Research* 9.4 (2020), pp. 502–510.

[25] Lue Lin, Luis Fernando D'Haro, and Rafael Banchs. "A Web-based Platform for Collection of Human-Chatbot Interactions." In: *Proceedings of the Fourth International Conference on Human Agent Interaction*. Biopolis Singapore, Oct. 4, 2016, pp. 363–366. DOI: 10.1145/2974804.2980500.

[26] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *nature* 521.7553 (2015), pp. 436–444.

[27] Ralf C. Staudemeyer and Eric Rothstein Morris. "Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks." In: *arXiv:1909.09586 [cs]* (Sept. 12, 2019).

[28] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." In: *2017 International Conference on Engineering and Technology (ICET)*. Aug. 2017, pp. 1–6. DOI: 10.1109/ICEngTechnol.2017.8308186.

[29] Rahul Chauhan, Kamal Kumar Ghanshala, and R.C Joshi. "Convolutional Neural Network (CNN) for Image Detection and Recognition." In: *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*. Dec. 2018, pp. 278–282. DOI: 10.1109/ICSCCC.2018.8703316.

[30] Richard Hahnloser et al. “Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit.” In: *Nature* 405 (July 2000), pp. 947–51. DOI: 10.1038/35016072.

[31] Abien Fred Agarap. “Deep Learning using Rectified Linear Units (ReLU).” In: *Computing Research Repository* abs/1803.08375 (2018).

[32] Ashish Vaswani et al. “Attention is All you Need.” In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[33] Yong Yu et al. “A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures.” In: *Neural Computation* 31.7 (July 1, 2019), pp. 1235–1270. DOI: 10.1162/neco_a_01199.

[34] Matthew E. Peters et al. “Deep Contextualized Word Representations.” In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237.

[35] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality.” In: *Advances in neural information processing systems* 26 (2013).

[36] Min Yang et al. “Sentiment analysis of Chinese text based on Elmo-RNN model.” In: *Journal of Physics: Conference Series*. Vol. 1748. 2. IOP Publishing. 2021, p. 022033.

[37] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *Computing Research Repository* abs/1810.04805 (2018).

[38] Alex Wang et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.” In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural*

Networks for NLP. Brussels, Belgium, Feb. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446.

[39] Pranav Rajpurkar et al. "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, Nov. 2016, pp. 2383–2392. DOI: 10.18653/v1/D16-1264.

[40] Rowan Zellers et al. "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 93–104. DOI: 10.18653/v1/D18-1009.

[41] Amil Merchant et al. "What Happens To BERT Embeddings During Fine-tuning?" In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online, Nov. 2020, pp. 33–44. DOI: 10.18653/v1/2020.blackboxnlp-1.4.

[42] Adina Williams, Nikita Nangia, and Samuel Bowman. "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana, June 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101.

[43] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. "Transformer models for text-based emotion detection: a review of BERT-based approaches." In: *Artificial Intelligence Review* 54.8 (Dec. 2021), pp. 5789–5829. DOI: 10.1007/s10462-021-09958-2.

[44] Victor Sanh et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *Computing Research Repository* abs/1910.01108 (2019).

[45] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *Computing Research Repository* abs/1907.11692 (2019).

[46] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, July 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703.

[47] Alec Radford et al. “Improving language understanding by generative pre-training.” In: (2018).

[48] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization.” In: *arXiv preprint arXiv:1412.6980* (2014).

[49] Dan Hendrycks and Kevin Gimpel. “Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units.” In: *Computing Research Repository* abs/1606.08415 (2016).

[50] Alec Radford et al. “Language models are unsupervised multitask learners.” In: *OpenAI blog* 1.8 (2019), p. 9.

[51] Yannick Le Cacheux, Hervé Le Borgne, and Michel Crucianu. “Zero-shot Learning with Deep Neural Networks for Object Recognition.” In: *Multi-faceted Deep Learning: Models and Data* (2021), pp. 127–150.

[52] Tom Brown et al. “Language Models are Few-Shot Learners.” In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[53] Robert Dale. “GPT-3: What’s it good for?” In: *Natural Language Engineering* 27.1 (Jan. 2021), pp. 113–118. DOI: 10.1017/S1351324920000601.

[54] Flor Miriam Del Arco et al. “Empathy and Distress Prediction using Transformer Multi-output Regression and Emotion Analysis with an Ensemble of Supervised and Zero-Shot Learning Models.” In: *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Dublin, Ireland, May 2022, pp. 239–244. DOI: 10.18653/v1/2022.wassa-1.23.

[55] Jochen Hartmann. *Emotion English DistilRoBERTa-base*. https : / / huggingface.co/j-hartmann/emotion-english-distilroberta-base/. 2022.

[56] Soujanya Poria et al. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, July 2019, pp. 527–536. DOI: 10.18653/v1/P19-1050.

[57] Myrthe Lammerse et al. "Human vs. GPT-3: The challenges of extracting emotions from child responses." In: *2022 14th International Conference on Quality of Multimedia Experience (QoMEX)*. 2022, pp. 1–4.

[58] Outi Bat-El. *CHILDES Hebrew Bat-El Corpus*. 2015.

[59] Paul Fletcher. *CHILDES English Fletcher Corpus*. Type: dataset. 2004. DOI: 10.21415/T51P55. URL: https://childes.talkbank.org/access/Eng-UK/Fletcher.html (visited on 04/09/2023).

[60] Hao Zhou et al. "Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 25, 2018).

[61] *Gated Recurrent Unit | Introduction to Gated Recurrent Unit(GRU)*. Analytics Vidhya. Mar. 17, 2021. URL: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-gated-recurrent-unit-gru/.

[62] Shimin Li, Hang Yan, and Xipeng Qiu. "Contrast and Generation Make BART a Good Dialogue Emotion Recognizer." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.10 (June 2022), pp. 11002–11010. DOI: 10.1609/aaai.v36i10.21348.

[63] Jingye Li et al. "HiTrans: A Transformer-Based Context- and Speaker-Sensitive Model for Emotion Detection in Conversations." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online), Dec. 2020, pp. 4190–4200. DOI: 10.18653/v1/2020.coling-main.370.

[64] Shalini Kapoor and Tarun Kumar. "Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network." In: *Multimedia Tools and Applications* 81.21 (2022), pp. 31107–31128.

[65] Mustaqeem, Muhammad Sajjad, and Soonil Kwon. "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM." In: *IEEE Access* 8 (2020). Conference Name: IEEE Access, pp. 79861–79875. DOI: 10.1109/ACCESS.2020.2990405.

[66] Nhat Truong Pham et al. "Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network." In: *Journal of Information and Telecommunication* 0.0 (2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/24751839.2023.2187278, pp. 1–19. DOI: 10.1080/24751839.2023.2187278. URL: https://doi.org/10.1080/24751839.2023.2187278.

[67] Elvis Saravia et al. "CARER: Contextualized Affect Representations for Emotion Recognition." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, Oct. 2018, pp. 3687–3697. DOI: 10.18653/v1/D18-1404.

[68] Saif Mohammad et al. "Semeval-2018 task 1: Affect in tweets." In: *Proceedings of the 12th international workshop on semantic evaluation*. 2018, pp. 1–17.

[69] Simeng Gu et al. "A Model for Basic Emotions Using Observations of Behavior in Drosophila." In: *Frontiers in Psychology* 10 (2019).

[70] *multi_nli · Datasets at Hugging Face*. Nov. 16, 2022. URL: https://huggingface.co/datasets/multi_nli (visited on 04/09/2023).

[71] Karen Gasper, Lauren A. Spencer, and Danfei Hu. "Does Neutral Affect Exist? How Challenging Three Beliefs About Neutral Affect Can Advance Affective Research." In: *Frontiers in Psychology* 10 (2019).

[72] David Dohan et al. *Language Model Cascades*. 2022. arXiv: 2207.10342 [cs.CL].