Text-based emotion detection to create a virtual avatar interview-training program.

Myrthe Lammerse

Thesis submitted for the degree of
Master in Informatics: Language Technology
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

Spring 2022

Text-based emotion detection to create a virtual avatar interview-training program.

Myrthe Lammerse

Text-based emotion detection to create a virtual avatar interview-training program.

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

Training police officers to interview abused children is expensive, challenging, and time-consuming. Unfortunately, interviewing abused children is a challenging task, as the majority of sexually abused children exhibit no visible signs of assault (Adams, Farst and Kellogg, 2018). Machine learning enables us to make the essential training schedule more accessible. We aim to design a training system aided by machine learning that can support the interview training with an interactive child avatar capable of meaningful interaction with the trainees. This thesis starts by focussing on the exploration of a language framework. It illustrates that non-fine-tuned GPT-2 models are ineffective in establishing a child-interview setting and cannot mimic an abused child. In addition, we create a sentiment pipeline within the RASA framework to extract emotions from sentences. We investigate different approaches to obtain the correct classifications. Both different models, as well as different data fine-tuning techniques, are tested and evaluated. In order to evaluate the approaches, we conducted three different user studies where users had to classify different transcripts excerpts as one of the possible predefined emotions.

Acknowledgements

I would like to thank my main supervisors Pål Halvorsen and Michael Alexander Riegler, as well as my informal supervisors Syed Zohaib Hassan and Saeed Shafiee Sabet. Without their consistent guidance and valuable feedback, this thesis would not have been the same. My appreciation also goes out to the faculty and staff of SimulaMet, who have given me the opportunity to contribute to the Talking Child-Avatar project. My recognition goes out to all researchers working on this project to make it a success. I would also like to thank the faculty and staff of the Department of Informatics at the University of Oslo. They have contributed to making the master's program a delightful and enlightening experience.

At last, I would like to express my gratitude to my friends and family. I am forever grateful for your loving support. Thank you for all your valuable advice, your check-ups, and for always believing in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Children who are subjects of abuse are prone to have cognitive, behavioural, and social problems. Additionally, they are susceptible to substance abuse, severe mental health problems, and death (Widom, 2014). To tackle these consequences, it is of great importance that child protective services (CPS) and law enforcement personnel interview the children following the existing guidelines (Poole and Michael E. Lamb, 1998) (Michael E. Lamb, 2016). Only by following these guidelines will interviewers be able to objectively interview the child to discover what happened, find the perpetrator, and ensure that this can be dealt with in court. These interviewers play an essential role in the process of protecting these children. Since the children are often both victims and key witnesses to the abusive incident, the informative interviews with child complainants play a crucial role in the investigation of the cases. (Michael E Lamb et al., 2011). Current research underlines the importance of interviewers' ability to follow empirically-based interview guidelines. Following this, there is a need for an efficient and cost-effective training module to prepare interviewers for real-life interview situations.

1.2 Problem Statement

The goal of the avatar research project is to design this training module that will prepare the CPS workers, law enforcement personnel, and police officers to interview children in cases of child maltreatment. The training module will be called the Talking Child-Avatar throughout this thesis.

The proposed multimodal avatar model (Baugerud et al., 2021) consists of several orthogonal parts; a language, an auditory, a sentiment, and a visual element. Out of these four main components, this thesis focuses on the sentiment part of the Talking Child-Avatar. More precisely, we wish to determine whether we can extract emotions from purely textual data and incorporate this into the system. The sentiment component classifies a sentence, or multiple sentences, as one emotion from a preset subclass of emotions and then returns a score belonging to each emotion. As shown

in Figure 1.1, the emotional component is an integral part of the system. It works together with both the visual and the audio part of the system as these outputs change based on the emotion that needs to be portrayed. In the figure, this part is indicated with the colour green. The sentiment classification component can both classify the input from the interviewer and the chatbot's generated response. The focus is on analysing the chatbot response, as the visual and audio output will utilise this for their output. The emotions will change these outputs by, for example, changing the facial expressions and the pitch of the speech.

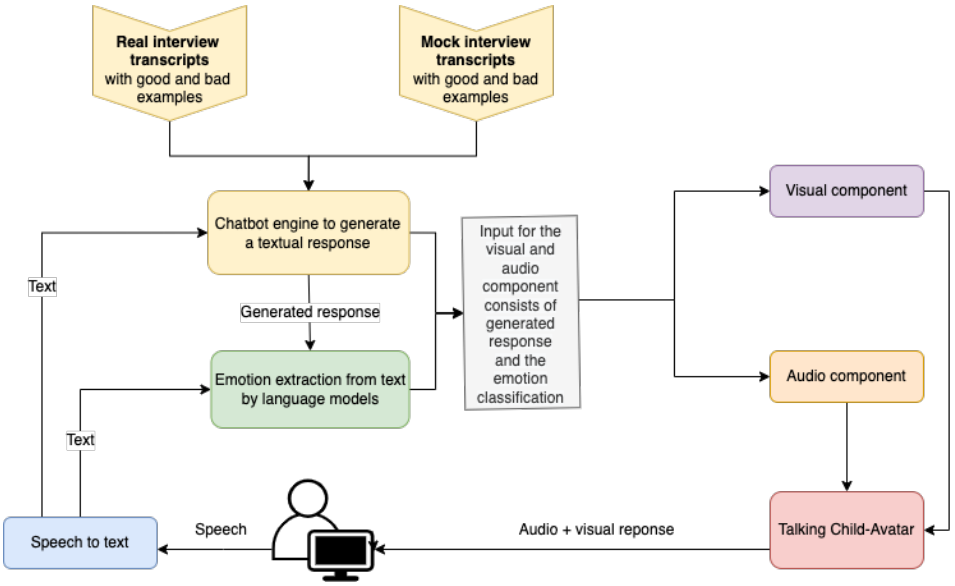

Figure 1.1: An overview of the Talking Child-Avatar architecture.

1.3 Research questions

This master thesis' primary aim is to extract emotions from child-police mock-interview transcripts to be used in different components of the Talking Child-Avatar. In addition, a secondary aim is to investigate which available framework is the best for creating a chatbot model that utilises the extracted emotions. Throughout this research thesis, the following questions will be answered with the aim of creating a solid foundation for the Talking Child-Avatar.

- What chatbot frameworks and language models are currently available? What are their pros and cons when it comes to this research project?
- How well can state-of-the-art models extract emotions from available transcripts compared to human annotation?
- How can we improve this emotion classification?

1.4 SimulaMet

This project was carried out in collaboration with Simula Metropolitan (SimulaMet). SimulaMet is a research division jointly owned by Simula Research Laboratory and Oslo Metropolitan University. SimulaMet specialises in activities on networks and communications, machine learning, and IT management.

1.5 Scope and limitations

The general aim and purpose of this research is to investigate how well start-of-the-art models can extract emotions. Moreover, it gives an overview of the available chatbot frameworks and language models in the context of the Talking Child-Avatar project. This study is limited to the ready-made available transcripts we obtained. There are no annotations available to accompany these transcriptions, nor are there resources to create them. Therefore, questionnaires have been conducted to get a general overview of the human extraction of emotions. These user studies took place between February and April of 2022. One limitation of this research is the insufficient sample size of the user studies conducted, nor were the participants experts in, i.e., the field of child psychology. Furthermore, this study does not cover the extraction of all possible emotions but merely focuses on the classification of a predefined subset. Another limitation to the user studies is that it can be difficult for humans to extract emotions from a single sentence or an excerpt due to the way the questions are set up. It is not possible to choose multiple emotions or 'none of the above'. This choice was made since the model always outputs the most probable emotion as well.

Moreover, most chatbot frameworks used in this study are very limited in scope as they are based on predefined rules and run on handcrafted rules to solve these problems. While the integration of advanced AI and machine learning has made considerable advances over the last few years, it has not yet reached its full potential. These technologies are better at handling multiple tasks compared to rule-based chatbots but have many limitations when it comes to long engaging conversations and the setting of context. These are both points of consideration when it comes to choosing the best chatbot framework, or language model, for our project. Since we lack annotated data, it is impossible to train a new chatbot model from scratch. Therefore, we have to use the pre-trained models that are openly available. Most available models are trained on data from adults instead of children, which may result in a deviation in language. In addition, the use of language models has consequences when it comes to handling sequences of variable lengths or an extensive vocabulary. It may result in a slow system while handling an extensive vocabulary when it matches each word of a sentence to the collection. Another limitation is that not all language models and chatbot frameworks available are freely available, and, therefore, we had to choose the best of the ones that were open source.

1.6 Research methods

Before starting the development of the chatbot, it is crucial to have a deep understanding of the technology and its scope. The Association of Computing Machinery (ACM) created guidelines (Comer et al., 1989) to assist in the execution of research. The aim is to help conduct research more coherently and consistently in the field of Computer Science. The three main paradigms presented are *theory, abstraction, and design*. The idea behind these paradigms is to provide a context for the computing discipline. The theory paradigm consists of four steps:

1. Characterise objects of study (definition)
2. Hypothesise possible relations among them (theorem)
3. Determine whether the relationships are true (proof)
4. Interpret results

The first part of the study is carried out in the form of a literature study based on this theory paradigm. The review of literature gives an overview of the accessible models and frameworks. This background information makes it possible to decide on the most suitable architecture with the required tools for the Talking Child-Avatar.

The implementation of the chatbot framework, the creation of the sentiment pipeline, and the sentiment extraction are based on the abstraction and design paradigms. Both the abstraction and the design paradigm also consist of four stages. The four stages of the abstraction paradigm are:

1. Form a hypothesis
2. Construct a model and make a prediction
3. Design an experiment and collect data
4. Analyse results

The design paradigm consists of the following stages:

1. State requirements
2. State specifications
3. Design and implement the system
4. Test the system

The results are essential for development-based research since they assist in understanding the process and the progress of the product development. As we need an intelligent chatbot system to conduct interviews with potentially abused children, we need to establish evaluation criteria to evaluate the results and track the progress of the chatbot system. The chatbot framework is evaluated by observation of its use of natural language. The sentiment analysis component is being evaluated by comparing the model output to human annotations.

1.7 Main contributions

This thesis exposes the current performance of language models and chatbot frameworks related, on the one hand, to an interview environment of mimicking an abused, neglected, and traumatised child and sentiment analysis on the other.

We show that a non-fine-tuned approach is not satisfactory when it comes to the pre-trained GPT-2 and DialoGPT models. The models are not able to recreate the child-interviewer setting that we require. Therefore, we developed an initial chatbot using the RASA framework. Not only is this a working example of Talking Child-Avatar, but we also included a sentiment analysis pipeline. Due to the closed-off RASA environment, we present another way to set up a RASA environment and connect to it using HTTP requests. This pipeline becomes the foundation for the Talking Child-Avatar as it is a tool to extract the emotion of both the user input and the bot's response. Other system components will use the emotion classification output as their input. The audio and visual parts depend on the emotions to alter their output to portray the correct characteristics.

We conducted multiple user studies to see how well humans can extract emotions from mock interviews. These user studies show excerpts of the mock-interview transcripts. Every participant has to classify an excerpt as one of the predefined emotions. The extraction gets compared between single-sentence excerpts and story excerpts with different window sizes. The windows are either of size 3, 5, or 7. We also argue why emotion extraction based on the whole story does not work and why context is essential when it comes to sentiment analysis.

We introduce and study multiple ready-available language models to use for sentiment extraction. We exploit how well these models work in a pure textual setting without the help of video material to substantiate the decisions. The models are implemented using two different settings. Firstly, the models are programmed in combination with a cosine similarity approach. Secondly, the models are exploited within the Huggingface zero-shot pipeline.

1.8 Thesis Outline

The thesis is structured as follows:

Chapter 2 - Background: This chapter starts with discussing the background material concerning the importance of interview training. Additionally, the concepts of Natural Language Processing, language models, and chatbot frameworks are addressed and explained in detail. The chapter resumes by giving some background on the difference between child and adult language and continues with background information on sentiment analysis. Then, the implementation of sentiment analysis is discussed. The chapter ends with highlighting some similar projects.

Chapter 3 - Chatbot Framework: This chapter starts with an exploitation of two different GPT-2 models concerning the chatbot framework. We continue with the development of a chatbot based on the RASA framework. Furthermore, the creation of a sentiment pipeline within the RASA framework is discussed.

Chapter 4 - Sentiment: The sentiment extraction models and strategies are examined in more detail in this chapter. We highlight four models from Chapter 2 and explain the methodology of the experiments carried out.

Chapter 5 - User Studies: Three different user studies have been conducted to observe the human annotation of emotions. This chapter provides more background information on these studies and reports the related results and observations.

Chapter 6 - Results and evaluation: In this chapter, the results are documented and explained. A comparison between the sentiment extraction of different models and the human annotation is made. The results of the different approaches are separated into multiple subsections. Furthermore, we highlight a couple of excerpts from the user studies conducted.

Chapter 7 - Discussion: This chapter discusses the results and limitations from Chapter 6.

Chapter 8 - Conclusion and future work: Finally, the findings of this thesis are concluded, and future work is suggested to improve the sentiment component of the Talking Child-Avatar.

Appendix A - user studies: The user studies conducted can be found in Appendix A. It shows the questions asked in combination with the possible answers. The Appendix is divided into three subsections for user study 1, 2, and 3 respectively.

Appendix B - user studies results: The answer distributions of the user studies from Appendix A can be found in this Appendix. It is divided into three different subsections corresponding to the three different user studies conducted, similar to Appendix A.

In summary, in this thesis, we both examine chatbot framework models as well as sentiment classification approaches. We created a sentiment classification pipeline in RASA. Furthermore, we conducted user studies that have been used to study how humans extract emotions and to evaluate the performance of the chosen strategies.

All the project software that has been made in accordance with this thesis is available on GitHub [1].

[1] https://github.com/MyrtheLammerse/thesis

Chapter 2

Background

This chapter contains the needed background knowledge and related work needed as prior knowledge for the rest of the research. The background consists of information about the societal problems of child maltreatment, language models, chatbot frameworks and sentiment analysis.

2.1 The societal problems of child maltreatment

According to The World Health Organization, child maltreatment is "the abuse and neglect that happens to children under the age of 18. It includes all types of physical and emotional ill-treatment, sexual abuse, neglect, negligence, and commercial or other exploitation" (*Child maltreatment* 2020). In 2019 alone, the Norwegian Child Protective Services (CPS) received 57.988 allegations of child maltreatment (Dyrhaug and Grebstad, 2020). During the COVID pandemic, which started in 2019 and caused whole countries to go into lockdown, there was a significant increase in the number of hospitalisations due to physical abuse among children between the age of 0 and 5 (Loiseau, 2021). The effects of child maltreatment are severe for the children involved, and the consequences are also felt within society. Currie and Spatz Widom (2010) indicate that adults with a childhood of either abuse, neglect, or both, have lower levels of education, employment, earnings, and fewer assets as adults. Furthermore, there is a 14% gap in the employment rate during middle age between individuals with histories of maltreatment and those without (Currie and Spatz Widom, 2010). In addition, Widom (2014) shows that child maltreatment can cause cognitive impairments, behavioural and societal problems, substance abuse, delinquency, disturbances in neurological development, severe mental health problems, and even death. To prevent these problems, it is of the highest priority to prosecute the offenders and, simultaneously, to ensure that innocent adults are not convicted of those criminal acts. Ensuring that these goals will be achieved is the only way we can protect vulnerable children from abuse and maltreatment.

2.2 The importance of interview training

The interview process should be impartial and fair to guarantee the prosecution of the guilty, as children are often both the victims and the principal witnesses in cases of child maltreatment. Medical confirmation is usually lacking, either due to lengthy reporting delays or, as is often the case, because penetration has not occurred (Adams, Farst and Kellogg, 2018). These circumstances lead to cases that are solely based on the statements given by those children. Due to the lack of supporting evidence, as the children often do not portray any visible signs of assault, interviews conducted by CPS and the police often play a crucial role in investigating abuse. These interviews are often the key to determining whether the suspect is convicted or not (Westcott, Davies and R. H. Bull, 2002). The quality of the investigative interview influences the child's credibility (Cassidy, Akehurst and Cherryman, 2020). The accuracy, competency, reliability, and truthfulness of the answers the children provide in their testimonies are affected by the interview quality. The research presented in Poole and Michael E. Lamb (1998) states that interviewers affect children by their choice of the physical environment for conducting interviews, their demeanour and behaviour, and their selection of questioning strategies, including their language. Guidelines and protocols have been constructed to ensure that these circumstances do not negatively influence the outcome of the interview.

The available training guidelines recommend that the interviewer uses open questions while avoiding option posing and suggestive techniques. The techniques to be avoided are the opposite; thus, using closed-ended questions and helping children form an answer (Poole and Michael E. Lamb, 1998) (Michael E. Lamb, 2016). The not-recommended techniques can sometimes negatively affect children's memory, resulting in inaccurate descriptions of the events that have occurred and may even result in false accusations (Goodman and Melinder, 2007). Despite the fact that interviewers act according to the best of their beliefs, it may not always lead to the desired result. People tend to be biased in favour of information that confirms their prior assumptions and rejects information that contradicts them. This is also known as the confirmation bias (Nickerson, 1998). Even professionals who are aware of this phenomenon, like scientists and doctors, are influenced by it. Therefore, interviewers from CPS and law enforcement interviewers should also be careful not to be affected by the confirmation bias. The confirmation bias, in combination with the efforts of police officers to find the truth, can lead to alarming situations. Beliefs and ideas that have already been established are difficult to change, even if contradictory evidence is available (S. E. Gorman and J. M. Gorman, 2016). Police interviewers must be careful that their confirmation bias does not result in faulty interview practices that can pose a severe threat to both the innocent suspect and the child. Not only should they be careful about this but they should also train themselves to battle these flawed acts.

New police officers often learn their training skills through training by observation. Training by observation means that the new police officers

observe experienced interviewers and learn the skills by watching. This way of training was the norm for a long time. Given that studies (Memon and Vrij, 2003; Irving, 1980; Redlich and Meissner, 2009; Cederborg et al., 2000; M. Johnson et al., 2015; R. Bull and B. Milne, 2004) have shown that most police officers who have not been adequately trained are poor interviewers; observation training will lead to the acquisition of bad practices. The reasons for the poor interview skills range from using the wrong interrogation technique (Memon and Vrij, 2003) to telling the suspect that it was in their best interest to confess (Irving, 1980). Therefore, observing an interviewer with insufficient skills will result in more interviewers who will be underperforming due to either having the wrong knowledge or having learned the inaccurate techniques. However, observation training is the cheapest method of training, as there are no extra costs for trainers and additional training materials. Multiple training modules have been developed to train the interviewers, but those modules are more effective when they are being used recurrently compared to a one-time use (Simon, Sousa and MacBride, 1997) (Pompedda, Zappalà and Santtila, 2015). The modules themselves do an excellent job of training police officers, but they can be very costly for different police departments, as they need to be used repeatedly. Not only do they need to be used repeatedly, but different ones need to be bought in order to subject the interviewers to new material every time. Consequently, departments need to pay a trainer, next to the new material, to teach interviewers the correct techniques for every training session. Unfortunately, due to the considerable monetary and personnel costs of formal interrogation workshops, 91% of police officers still learn interrogation skills from their, often poorly trained, colleagues (Cleary and Warner, 2016).

This research project aims to develop a Talking Child-Avatar that is both cost-effective and efficient in providing interview training. The aim is to generate a system that can produce its own storylines so that it can be used repeatedly. CPS workers, law enforcement personnel, police officers, and others who require interview skills in a setting where child maltreatment is present can use the system to train themselves. They can reuse this chatbot as often as necessary, resulting in effective training results. However, in contrast to other training modules, this will be less expensive since there is no need to hire trainers for every teaching session. Not only do they cut costs by not hiring trainers, but the aim is also to let the chatbot create its own training modules so that only one training module needs to be acquired.

In this thesis, we will focus on the chatbot framework and the sentiment component of the Talking-Child Avatar. It is of utmost importance to have the avatar portray the correct emotions due to the existence of emotional intelligence and the corresponding training. Emotional intelligence is the ability to perceive, use, understand, and manage emotions (Salovey and Sluyter, 1997). The level of emotional intelligence that an investigative interviewer portrays influences the execution and performance of the emotional labour (Joseph and Newman, 2010). Emotional labour refers to jobs where the employees are expected to be alert to emotions and act accord-

ingly. These jobs have guidelines and rules in place to handle this (Hochschild, 2012). The results of an investigative police interview concerning adults depend on how officers handle the interviewee's emotions. The outcome, the interviewees' well-being, and therapeutic jurisprudence are positively influenced if the interview is conducted in an emotionally intelligent, appropriate way (Risan, Binder and R. J. Milne, 2016). Research addressing emotional intelligence in child interviews is limited but one recent study found support for the equivalent mechanisms among CPS workers and mental care service psychologists addressing abuse in interviews with children (Albaek, Kinn and Milde, 2018). Therefore, the training module must effectively express the right emotions to train CPS workers and police officers within the investigative interviewing field. The emotions can be used in both the visual and the auditory outputs that is, by changing, for example, the facial expressions and the speed of the speech.

Research shows that there are seven universal emotions based on facial cues (Ekman and Friesen, 1986; Ekman and Heider, 1988; Matsumoto, 1992). These seven emotions are *enjoyment, sadness, anger, disgust, contempt, fear, and surprise*. Other research defines the four basic emotions as a smaller subset of these universal emotions (Gu et al., 2019). The four basic emotions are *enjoyment, sadness, anger, and fear*. The goal is to portray these sets of emotions during the conversation with different degrees of expression based on different personas. The different degrees of expression are necessary and can be created by developing different personas that express emotions to different degrees. This is important since the expression of emotions can vary considerably between children, especially traumatised children. Traumatised children do not always follow the general rules of emotional expression. They can numb all their emotions or show positive and negative emotions in a context that is not considered normal (Kerig et al., 2016).

2.3 Natural Language Processing and its components

In order to develop the chatbot, we need to start with explaining what Natural Language Processing (NLP) and its components Natural Language Understanding (NLU) and Natural Language Generation (NLG) are. NLP is the process of converting human language into structured, machine-readable, and understandable data. NLU is the task of extracting the meaning of a sentence using semantic and syntactic techniques. Lastly, NLG is the task of generating language from scratch or based on a given input (Kumar, 2018) (Kavlakoglu, 2020).

The functioning of a chatbot is based on its NLU component. In theory, NLU is a sub-field of computer science where the focus lies on learning and understanding human language (Hirschberg and Manning, 2015). In practice, NLU uses machine learning and other NLP techniques to extract structured information from unstructured user input. However, for this project, the NLG part is just as important since the goal is to create a chatbot that is able to understand the input that a human gives and then generates

its own response. This response follows a storyline that the chatbot comes up with. The generation part is essential to engender new narratives to ensure the interview trainees are subjected to new material every time they use the Talking Child-Avatar.

2.4 Language Models

Whenever you read about the progress that is being made with state-of-the-art language models, you would believe that we can use those models to write theses, articles, and anything that you can imagine. However, is this really the place we are at right now? And how did we arrive at the point that we have reached so far? This section will discuss both the current state of language models as well as describe some available models in more detail.

2.4.1 An analysis of language models

Theoretically, Language Models (LMs) are able to generate any sentence that humans can. LMs can do so by calculating the probability of any possible string. We can classify the language models in the NLG part of NLP. One of the first tasks that LMs attempted to decipher was the task of completing sentences and texts. When the SWAG dataset (Zellers, Bisk et al., 2018) was first presented, it showed that while humans could solve the resulting inference problems with a high accuracy of 88%, various state-of-the-art language models at that time struggled with this task. The accuracy obtained by these models was less than 60% (Zellers, Bisk et al., 2018). This poor performance was first thought to have been caused by the absence of commonsense reasoning within the models. It was believed that it was necessary to have this to solve this natural language inference task. However, the models are becoming better, and results have improved on a variety of tasks, from inference to sentiment analysis assessments. According to Hirschberg and Manning (2015), there are four key components that have and continue to contribute to the advancement of natural language processing, generating, and understanding. These four are:

- An immense increase in computing power.
- The availability of vast amounts of linguistic data, mostly due to the internet.
- The development of successful machine learning methods.
- The understanding and structure of human language that we have obtained through research.

The training of every language model follows a specific strategy. As of now, there are three existing strategies for using pre-trained language models on downstream tasks. Downstream tasks are supervised tasks

that are used to measure how well the pre-trained model performs. The first strategy to solve these tasks is based on features. A feature-based approach uses the task-specific parameters from the downstream task and combines those with the pre-trained representations from the language model. The other strategy is called the fine-tuning policy. This strategy uses a minimal amount of task-specific information. To be able to carry out the downstream task, it fine-tunes the pre-trained parameters. The final one is a zero-shot strategy in which a model is trained on some data but gets assessed on a completely different set of data. Next to a learning strategy, every LM also has an architecture and a direction of self-attention. The architecture of a neural network tells us how new information flows through the system. The architecture of a recurrent neural network (RNN) can be seen in Figure 2.1. There is an input x_t and an output h_t based on the input. The network loop allows information to be passed from one network step to the next. These feedback loops in the recurrent layer allow the network to have some memory. However, this memory is often insufficient when a problem requires a long-term memory structure.

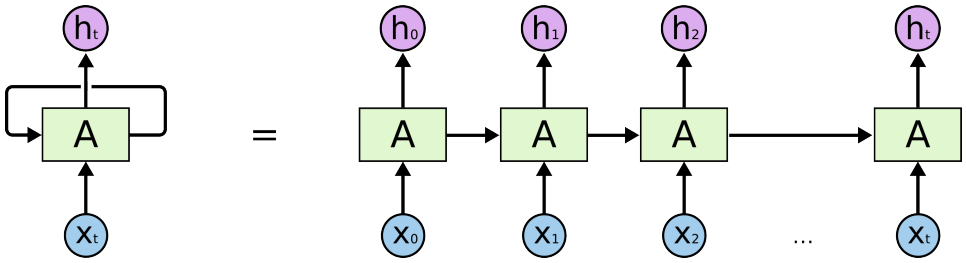

Figure 2.1: The architecture of an RNN (Olah, 2015).

.

Long Short Term Memory (LSTM) networks are a variation of the RNN network but are, on the contrary, capable of long-term dependencies. Similar to RNNs, LSTMs also have a chain-like structure but every link has four layers, compared to only a single neural network layer in an RNN. The difference can be seen in Figures 2.2 and 2.3. The yellow boxes represent the different activation functions within these networks. These extra layers introduce the ability to introduce extra inputs or forget information, and therefore, it is better to preserve the long-range dependencies. Similar to RNNs and LSTMs, transformers are designed to process sequential input data such as natural language. However, unlike RNNs, transformers do not necessarily process data sequentially.

The direction tells us about the architecture and how previous tokens can be seen in the self-attention layers. A self-attention mechanism allows the input to interact with every other input but also with itself. A uni-directional self-attention mechanism can only see the tokens from left-to-right, whereas a bidirectional self-attention mechanism gives us both left and right contexts. A deeply bidirectional representation uses a Masked Language Model (MLM) pre-training objective. An MLM masks out some randomly chosen part of the input, and then the goal of the objective is to predict the original word based solely on the left-over context that is

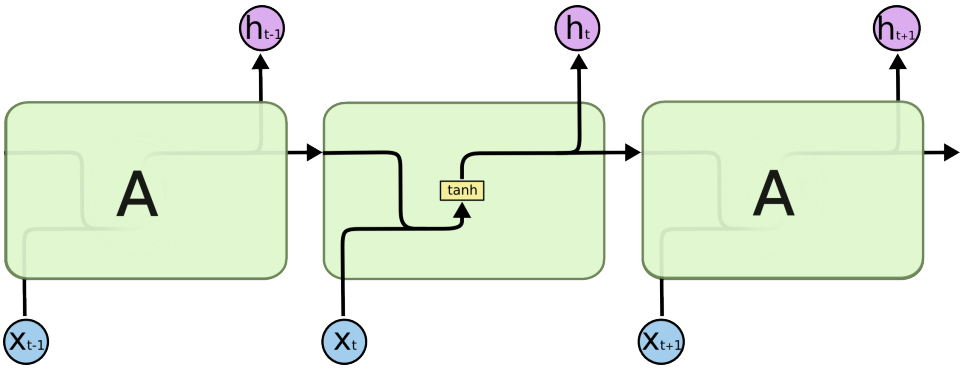

Figure 2.2: The neural network layer in an RNN network (Olah, 2015).

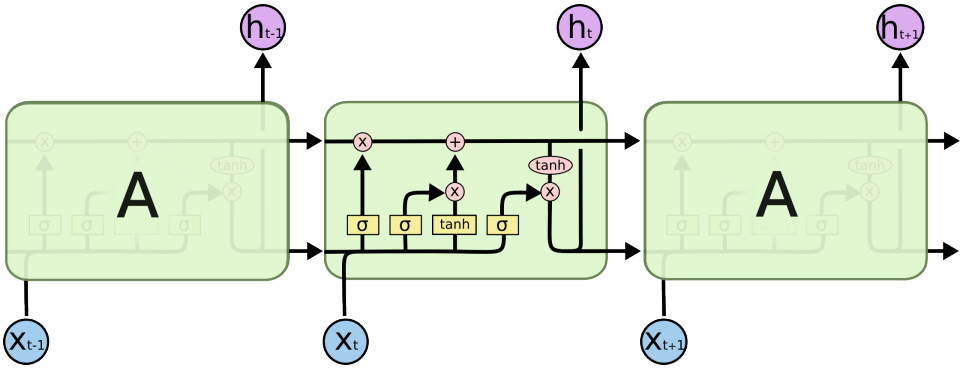

Figure 2.3: The neural network layers in an LSTM (Olah, 2015).

still present. In contrast to traditional left-to-right RNNs, the MLM objective makes it possible for language models to combine representations from both directions. Traditional RNNs see the words from left-to-right, and autoregressive, meaning that it predicts future values based on past values, models mask the future models internally, which makes MLM stand out as it learns a deeply bidirectional representation of the sentence. The difference between bidirectional and deeply bidirectional is due to the way the bidirectional representations are created and combined.

A couple of the most influential LMs will be described in the rest of this section. This is essential as we may use them for the Talking Child-Avatar. On the one hand, the LMs can be used as a technique to simulate the child by letting it generate language and storylines. On the other hand, LMs can be used to extract emotions from textual data. An overview of these models can be found in Table 2.1.

2.4.2 Language models in context of the Talking Child-Avatar

Language models are often evaluated according to the GLUE benchmark (Wang et al., 2018). The GLUE benchmark is a collection of different tasks that can be used to measure the performance of a model across a set of various Natural Language Understanding problems. GLUE consists

Language Model	Number of parameters	Architecture	Direction of self-attention
BERTbase	110 million	Transformer	Deeply bidirectional
BERTlarge	340 million	Transformer	Deeply bidirectional
CTRL	1.63 billion	Transformer	Bidirectional
ELMo small	13.6 million	BiLSTM	Bidirectional
ELMo medium	28 million	BiLSTM	Bidirectional
ELMo original	93.6 million	BiLSTM	Bidirectional
GPT-2	1.5 billion	Transformer	Unidirectional
GPT-3	175 billion	Transformer	Unidirectional
Grover base	124 million	Transformer	Unidirectional
Grover mega	1.5 billion	Transformer	Unidirectional
HAIM large	345 million	Transformer	Bidirectional
Haim-1.5	1.5 billion	Transformer	Bidirectional
RoBERTa	110 million	Transformer	Deeply bidirectional
SenseBERT base	133 million	Transformer	Deeply bidirectional
SenseBERT large	370 million	Transformer	Deeply bidirectional
DeBERTa large	1.5 billion	Transformer	Deeply bidirectional
DistilBERT large	66 million	Transformer	Deeply bidirectional
BART large	110 million	Transformer	Deeply bidirectional

Table 2.1: An overview of the language models mentioned in Section 2.4.

of nine English sentence comprehension tasks covering many domains, quantities of data, and difficulties. There are two single-sentence tasks, three similarity and paraphrase tasks, and four inference tasks. One of the single-sentence tasks is SST-2, which stands for The Stanford Sentiment Treebank (Socher et al., 2013). It consists of sentences from movie reviews and human annotations of their sentiment. The corresponding task is to predict the sentiment of the given sentiment. The result of this task can be an important indication for the sentiment classification of the Talking Child-Avatar. However, it is essential to keep in mind that this is a classification of either the positive or the negative class and not on the subset of emotions that we would like to extract. For the chatbot part of the thesis, the inference results can be an important indicator. The highest performing models come close to human results on the GLUE tasks. These results are promising for the Talking Child-Avatar. However, it is vital to keep in mind that none of these models are trained in the specific child-interview setting that we want to create. They are also not fine-tuned on child language but merrily on adult language.

Not every model has reported its scores on the GLUE benchmark, and, therefore, it is not an option to pick the highest-performing models. However, based on the characteristics of the models, we can still make an educated choice. It is important to experiment with different models and evaluate the results to see what the best option is as it is not predetermined how each model will perform.

2.4.3 BERT

BERT is introduced in Devlin et al. (2018) and stands for Bidirectional Encoder Representations from Transformers. There are two crucial

steps in creating BERT: pre-training and fine-tuning. The bidirectional representation of BERT is enabled by using MLM. The other pre-training objective is *Next Sentence Prediction* (NSP). This training objective aims to predict sentence relationships. After the pre-training is done, the transformer only needs to be fine-tuned. Fine-tuning is extremely task-sensitive, and therefore we need task-specific inputs and outputs. Whenever those are ready, it is possible to plug those inputs and outputs into BERT and then fine-tune all the parameters end-to-end. BERT is what we call an autoencoding model since it corrupts the input tokens and then tries to reconstruct them to the original sequence.

2.4.4 ELMo

ELMo stands for Embeddings from Language Models and is presented in Peters et al. (2018). ELMo uses an unsupervised feature-based approach, meaning it uses both the task-specific and the pre-trained parameters. ELMo concatenates the extracted context-sensitive features from both a left-to-right language model and a right-to-left one. Since the language model considers both directions, it is considered a bidirectional language model. However, due to the concatenation, it is not considered a deeply bidirectional model, as BERT is. ELMo is an example of an autoregressive model. These models mask the entire sentence and can only use what it has previously seen in the text and not what is coming in the following parts. An autoencoder model, like BERT, does have access to this information.

2.4.5 Grover

Grover is affiliated with the Allen Institute for Artificial Intelligence and is introduced in Zellers, Holtzman et al. (2019). The objective of this language model is to fight against the spread of fake news on the internet. The language model is created in three different model sizes. The smallest, called Grover-base, has 12 layers and 124 million parameters. Grover-large consists of 24 layers and has a total of 355 million parameters. The largest model is called Grover-mega. This model has 48 layers and 1.5 billion parameters in total. The language model is pre-trained on the RealNews corpus (Zellers, Holtzman et al., 2019), a 120-gigabyte corpus consisting of news articles from Common Crawl[1], a non-profit organisation dedicated to providing a copy of the internet. The language model was then pre-trained on randomly sampled sequences from RealNews, with a length of 1024 tokens.

2.4.6 GPT-2

The researchers that created GPT-2 built their own webscraper. Using this webscraper, they scraped the web to create a dataset called *WebText* (Radford et al., 2018). WebText consists of the context of all the links that are mentioned on Reddit. Reddit is a social media platform where users

[1] https://commoncrawl.org/

can start conversations on pretty much every topic. However, not all links were used; they only used links that received at least 3 likes, or karma as those are called on Reddit. The content of those links was evaluated and cleaned up before it was added to their dataset. After Radford et al. (2018) had completed the creation of the dataset, they trained a model on this data. They ended up with a model called GPT-2, a 1.5 billion parameter Transformer that achieves outstanding results on 7 out of the 8 tested language modelling datasets in a zero-shot setting, which means that there was no training done from samples belonging to those datasets. GPT-2 is an autoregressive model, just like ELMo.

2.4.7 GPT-3

GPT-3 is an autoregressive model, just like its predecessor GPT-2 but achieves better results (Brown et al., 2020). GPT-3 can produce higher-quality text due to, among other things, its bigger embedding size and its wider context window size. After training, they ended up with a 175 billion-parameter language model. It follows the fine-tuning strategy of learning for downstream-task.

2.4.8 HAIM

The language models mentioned above are autonomous and are nearly impossible to control. HAIM, on the other hand, is what is called an interpolating language model. This means that HAIM generates text between a human-written beginning and a human-written ending. The length of the body can be specified to make the generated text more to your liking. To create HAIM, the same transformer-based architecture as GPT-2 and Grover has been used. The model is trained on *OpenWebText* (Radford et al., 2018), created by OpenAI. OpenAI is also the company behind the GPT language models. After training, HAIM consists of a total of 345 million parameters. HAIM is worth mentioning but is not a scaleable solution for the Talking Child-Avatar as we do not want to supply both an beginning- and ending prompt.

2.4.9 CTRL

The Conditional Transformer Language Model, or in short CTRL, is developed by Salesforce. CTRL is a language model consisting of 1.63 billion parameters (Keskar et al., 2019). The model has been trained on 140 gigabytes of text. The text has been obtained from a wide variety of domains, ranging from Wikipedia in different languages to data from the United Nations. The unique selling point of CTRL is that the language model is trained using control codes based on the desired style, content, and task-specific behaviour. These control codes can distinguish between the desired features of the generated texts.

2.4.10 SenseBERT

SenseBERT is an extension of the original BERT model. The original BERT model is solely self-supervised, what is achieved by MLM. SenseBERT combines the original pre-training strategies with weak supervision on word level. By using this weak-supervision approach it cannot only predict the mask words but also their meaning (Levine et al., 2020). To settle on the semantic meaning of masked words, they use WordNet (Fellbaum, 1998). On average, SenseBERT outperforms BERT on the GLUE Benchmark.

2.4.11 RoBERTa

Liu et al. (2019) introduces RoBERTa by altering the BERT training policy. The modifications that were introduced were the following:

- A longer training time combined with larger batches and more data.
- Removal of the next-sentence prediction objective.
- Longer sequences during training
- Dynamically changing the masking pattern.

By evaluating the effects of hyper-parameter tuning and the modification of the training set size, they were able to train a model that outperforms the original BERT model.

2.4.12 DeBERTa

DeBERTa[2] is an improvement on the BERT and RoBERTa models when it comes to a majority of NLU tasks. Improvements are made by the use of disentangled attention and an enchanced mask decoder. Disentangled attention means that each word is represented using two vectors which encode the contents and positions of the word, respectively, and a word's weights of attention are calculated from disentangled matrices on their contents and relative positions. An enhanced mask decoder is used to include absolute positions in the decoding layer to predict masked tokens during the pre-training of the model (He et al., 2021).

2.4.13 DistilBERT

DistilBERT[3] is a transformers model, similar to BERT since it is pre-trained on the same corpus in a self-supervised way but smaller and faster. More precisely, it was pre-trained with three objectives (Sanh et al., 2020):

- Distillation loss; keep the loss of knowledge as small as possible compared to the BERT base model by training it to return the same probabilities.

[2] https://huggingface.co/Narsil/deberta-large-mnli-zero-cls
[3] https://huggingface.co/typeform/distilbert-base-uncased-mnli

- MLM as explained in Section 2.4.3
- Cosine embedding loss; keep the loss given two tensors small. If this is succeeded, the model generates hidden states as close as possible to the BERT base model. The cosine loss is measured given two input tensors x_1 and x_2 and a tensor label y with values 1 or -1. The loss function is formally written as:

$$loss(x, y) = \begin{cases} 1 - cos(x_1, x_2) & if y = 1 \\ max(0, cos(x_1, x_2) - margin) & if y = -1 \end{cases}$$

Subsequent to training with these three objectives in mind, DistilBERT is able to learn the same internal representations about the English language compared to BERT, while at the same time being faster for downstream tasks.

2.4.14 BART

The BART Large MNLI model[4], trained by Facebook, is the newly saved BART Large model with a checkpoint right after being trained on the MultiNLI (MNLI) dataset (Williams, Nangia and Bowman, 2018). The original BART Large (BART) model (Lewis et al., 2019) is a sequence-to-sequence model pretrained on English data. It has a bidirectional encoder and a autoregressive decoder.

2.5 Ethical side of language models

Since the above-mentioned language models are capable of producing language in a way that is almost indistinguishable from human language, the question arises as to what ethical concerns emerge. Language models have been shown to be used to generate fake news that can influence elections (Faris et al., 2016) (Molina et al., 2019) by writing target propaganda or they can be used to earn money in an illegal way (Davey-Attlee and Soares, 2017). Even if language models are developed with the best intentions, there is no control over what other people may do with your model. It was also never the intention to create these models so that they could create fake new but is does happen quite often nowadays. Luckily, there is a way to combat this since the best way to combat language models that generate fake news is by using the same language models. The reason for this is that these language models are able to find statistical regularities while humans consider the text generated by a model also as human-like (Zellers, Holtzman et al., 2019). Zellers, Holtzman et al. (2019) writes that the best current discriminators can classify neural fake news from real human-written news with 73% accuracy, assuming access to a moderate level of training data. The Talking Child-Avatar project uses ideas from the fake news area in order to create the system. This may

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

be a bit controversial in itself and in many respects unethical, making it quite controversial. However, technology is used in a positive way in the context of this project. An favourable ethically aspect is that, by creating the possibility of interviewer training with virtual children, there is no need to practice interviewing skills on real, vulnerable children. The additional advantage is that there is no requirement for the informed consent of the avatar, an obligation when interviewing real-life children.

So even though language models seem like the solution to combat the misuse of language models, it is still important to keep ethical concerns in mind when creating a powerful language generating tool. Especially since we are working on such a delicate subject as child neglect and abuse. One of the seven principles of privacy by design introduced by Cavoukian (2009) is the following: *a process firmly inclusive of ethics at all stages and levels is less likely to create accidental harm*. Therefore, it is important to give ethics a prominent spot within this research project. One of the most important points is to pay attention to fairness, bias, and discrimination (Leidner and Plachouras, 2017). To combat discrimination and build a fair model, it is important to be aware of your training data. A model learns its decision based on the training data that it is given. Whenever this data includes biased human decisions or reflect historical or social inequities, it can result in an unfair system. This is especially important within law enforcement where equality is of highest importance but is also subjective to racism. Police officers have been in bad daylight due to racism, police brutality, and the combination of these two (Easton, 2009) (Pierson et al., 2017) (Khan, 2020) and that is not something we want to enforce with this project.

Not only is it important to develop a system that is not biased in any sense. The ethical concerns are also important because the system works with sensitive data and needs to be sensible to privacy concerns that arise with this. We want to preclude that the system will be used with the wrong intentions. It is still unclear and not yet decided to what degree the code and data of the Talking Child-Avatar will be freely available.

2.6 Research in chatbots frameworks

Conversational agents, or chatbots, are software programs that communicate with people by voice or text in natural language (Lucas et al., 2014) (Abdellatif, Badran, Costa et al., 2020). The use of chatbots has greatly increased over the last couple of years. Companies and governments start to see the potential of using chatbots as a means to transfer knowledge. One of the strengths of chatbot systems is that they can give a single conclusive answer to a question. The user of a chatbot is then less overwhelmed by a long list of information that may or may not be relevant ('Chatbots in the fight against the COVID-19 pandemic' 2020). Not only can it be more specific about the information it gives, as said before, it is also able to hold conversations. So, a chatbot can be used in a wide variety of circumstances. Therefore, it is possible to think that chatbots can also be used as a training mechanism where they converse with the trainees and give feedback after

such training session.

Before describing the various NLUs, it is important to know the meaning of both *intent* and *entity*. An intent addresses the planning between what a user says and what move ought to be made by the chatbot. An entity sets boundaries from natural language inputs. Any important information that you need to get from a user will be connected to a relating entity (Canonico and Russis, 2018).

In the next subsections, we will elaborate on a couple of start-of-the-art NLUs, or in other words, chatbot frameworks. Chatbot frameworks can be a solution to the lack of annotated data, as it often follows a predefined storyline compared, so it involves little to no training. A disadvantage, however, is that they cannot generate their own storyline. Furthermore, We will elaborate on the differences in performance between the NLUs in the next sections. However, it is important to remember that these experiments are influenced by both the genre of text and the time the research was carried out, since NLUs are getting better and smarter every day.

An overview of all chatbot frameworks can be found in Table 2.2.

2.6.1 RASA

RASA is an open-source dialogue system. The system can be used to build conversational systems, such as chatbots. RASA's architecture is modular by design. This allows easy integration with other systems that may be useful for specific projects (Bocklisch et al., 2017). There is one dialogue tracker per conversation session, and this is the only stateful component in the system, meaning that this is the object where the dialogue state is saved. The state of a conversation can be reconstructed by replaying the exact order of the previous events. RASA predicts which action to carry out based on a predefined list. The tool can be used both for intent classification as for entity extraction. The natural language understanding module is based on pre-defined pipelines. The pipeline that is recommended by RASA is spacy_sklearn. This pipeline makes use of GloVe vectors. Moreover, it is possible to change the GloVe vectors for custom domain-specific word embeddings. This can be a huge advantages since our training chatbot has a very specific domain wherein it will operate, but there is a need for annotated data to make it a success.

2.6.2 Gensim

On a similar note to RASA, Gensim is free and open-source. It is Python package for topic modelling, document indexing and similarity retrieval with large corpora. It can be used to create chatbot, but is not a chatbot framework on its own. It represents documents as semantic vectors within a vector space. It is unsupervised and thus only needs documents for training (Řehůřek and Sojka, 2010). The difference between Gensim and RASA is that Gensim transforms each document to a vector and then uses a model based on this, whereas RASA follows precise defined stories. It

depends on the amount and quality of the available data whether this has a positive or a negative influence.

2.6.3 Microsoft's LUIS

Language Understanding Intelligent Service, or in short LUIS, is owned by Microsoft (Microsoft, 2021a). The pricing is based on 1.000 requests and differ whether the requests are written or spoken word. For 1.000 requests you pay $1,50 per 1000 written messages and $5,50 per spoken ones. For text requests, the limit of a transaction is the query length of 500 characters. For speech requests, this limit lies at a 15 second long query (Microsoft, 2021b). However, since it is also part of Microsoft's Azure, it also possible to get it as a combination of both. Since it is part of Microsoft's cloud service, LUIS can be categorised as a cloud-based chatbot platform. According to Canonico and Russis (2018), it is possible to use existing models from Bing and Cortana, which are extensive, as they have been trained for a long time already.

2.6.4 Amazon's Lex

Similar to LUIS (Section 2.6.3 and IBM Watson(Section 2.6.7, Amazon's Lex is also part of the companies cloud service, called Amazon Web Services (AWS). Therefore, both services also share the pricing plan. It can be both used for voice and text requests. To be more exact, it provides functionalities within the field of automatic speech recognition (ASR) so that it can convert speech to text. The same deep learning technologies that are being exploited by Amazon's Alexa are also being used by Amazon Lex.

2.6.5 DialogFlow

DialogFlow was previously known as api.ai, until Google acquired the company behind the conversational tool (Huffman, 2016). DialogFlow is maintained by Google and the trial version is free to use. This entails that you get a limited quota of requests. Furthermore, the free version is only suitable for small to medium and simple to moderately complex conversational agents. The free trial version is, however, a suitable way to experiment with DialogFlow (Google, n.d.). Both the paid and the free version support various languages, different programming languages and are able to subtract the entities out of a conversation. The DialogFlow NLU extracts intents and entities from the user input by employing a custom made NLP model. DialogFlow has been used in research before. Abdellatif, Badran and Shihab (2020) motivated their choices because it is both a strong NLU model as well as that it can be easily integrated with 14 different platforms, and it supports more than 20 languages. Those languages can even be mixed within one single intent (Canonico and Russis, 2018). Furthermore, it is possible to use either the DialogFlow webinterface or the

API. In the research conducted by Gregori (2017), DialogFlow performed best out of wit.ai (Section 2.6.6), LUIS (Section 2.6.3)and Lex (Section 2.6.4).

2.6.6 Wit.ai

Where DialogFlow is maintained by Google, wit.ai is kept up by another big tech company, namely Facebook. Wit.ai is completely free to utilise, even for commercial use, support multiple languages, and also support three programming languages. Just like DialogFlow, LUIS, and Lex, wit.ai is also a cloud-based chatbot platform. The programming languages that wit.ai offers clients are Python, Ruby, Go, and Node.js (Liao et al., 2021). The difference between the wit.ai NLU and other NLUs mentioned is that it focuses on extracting meaning from a single sentence instead of from the whole conversation. This will probably be less suitable for our research, since the answers, and follow-up questions rely on what has previously been said.

2.6.7 IBM Watson

Since Watson is part of the IBM Bluemix cloud services, it also is included in its pricing plan. However, the free tier offers up to 10.000 API calls per month. Again, this also means that Watson is, just like LUIS, a cloud-based chatbot platform. Watson is trained on an impressive number of one billion words from Wikipedia. It tries to obtain as much context as possible to detect entities and intent. It retrieves this context from both the user's input and the corpus that is available to Watson. Watson was tested as the best NLU compared to DialogFlow (Section 2.6.5), Amazon's Lex (Section 2.6.4), and Microsoft's LUIS (Section 2.6.3) by Canonico and Russis (2018).

In the end, all frameworks have similarities and differences. Therefore, it is hard to choose the best option of all without actually testing them out within our own system in combination with our own data. At this moment, the criteria for picking one is that it can work well with a relatively small amount of annotated data and is still able to create the required interview setting.

2.7 Difference between child and adult language

Most chatbots and language models currently available are based on adult language. In order to be able to develop a system that mimics a child, it is important to be able to distinguish between the language that adults use and the language that children use. Vocabulary wise, humans peak between after they reached the age of 35 (Hartshorne and Germine, 2015). Not only do adults have a bigger vocabulary, they are also better able to answer general knowledge questions and better in explaining why things happen. The lexicon that the chatbot uses should thus not only be simpler but also the form of the answers should be grammatically different.

Chatbot framework	Supported languages	Supported programming languages / API	Prices
RASA	Every language as long as you have embeddings.	Python	Free
Gensim	Every language as long as you have embeddings.	Python, Cython	Free
Microsoft's LUIS	ar, zh-cn, nl, en-us, fr-ca, fr-fr, de, gu, hi, it, ja, ko, mr, pt-br, es-lx, es-sp, ta, te, tr	Cloud service	1,50 per 1000 written messages and $5,50 per spoken ones.
IBM Watson	ar, zh-cn, zh-tw, cs, nl, en-us, fr-fr, de, it, ja, ko, pt-br, es-sp	Java, C, or Python	Free up to 10.000 API calls per month
Amazon's Lex	de, en-au, en-gb, en-us, es-lx, es-sp, fr-ca, fr-fr, it, ja	Cloud service	$0.004 per speech request $0.00075 per text request
DialogFlow	bn, bn-bd, bn-in, zh-hk, zh-cn, zh-tw, da, nl, en, en-au, en-ca, en-gb, en-in, en-us, tl, tl-ph, fi, fr, fr-ca, fr-fr, de, hi, id, it, ja, ko, ms, mr, mr-in, no, pl, pt-br, pt-pt, ro, ro-ro, ru, si, si-lk, es, es-lx, es-sp, sv, ta-in, ta-lk, ta-my, ta-sg, te, te-in, th, tr, vi, vi-vn	Python	$0.007 per text request $0.06 per audio minute
Wit.ai	ar, bn, br, ca, zh-cn, nl, en, fi, fr, de, hi, id, it, ja, ka, ko, ms, ml, mr, pl, pt-pt, ru, si, es, sv, tg, ta, te, th, tr, ur, vi	Python, Ruby, Go and Node.js	Free

Table 2.2: An overview of the chatbot frameworks mentioned in Section 2.6.

Arabic (ar), Bengali (bn), Bengali - Bengladesh (bn-bd), Bengali - India (bn-in), Burmese (br), Catalan (ca), Chinese - Cantonese (zh-hk), Chinese - Simplified (zh-cn), Chinese - Traditional (zh-tw), Czech (cs), Danish (da), Dutch (nl), English (en), English - Australia (en-au), English - Canada (en-ca), English - Great-Brittain (en-gb), English - India (en-in), English - United States (en-us), Filipino (tl), Filipino – The Philippines (tl-ph), Finnish (fi), French (fr), French - Canada (fr-ca), French - France (fr-fr), German (de), Gujarti (gu), Hindi (hi), Indonesion (id), Italian (it), Japanese (ja), Kannada (ka), Korean (ko), Malay (ms), Malayalam (ml), Marathi (mr), Marathi – India (mr-in), Norway (no), Polish (pl), Portuguese - Brazil (pt-br), Portuguese – Portugal (pt-pt), Romanian (ro), Romanian – Romania (ro-ro), Russian (ru), Sinhala (si), Sinhala – Sri Lanka (si-lk), Spanish (es), Spanish – Latin America (es-lx), Spanish - Spain (es-sp), Swedish (sv), Tagalog (tg), Tamil (ta), Tamil – India (ta-in), Tamil – Sri Lanka (ta-lk), Tamil – Malaysia (ta-my), Tamil – Singapore (ta-sg), Telugu (te), Telugu – India (te-in), Thai (th), Turkish (tr), Ukrainian (uk), Urdu (ur), Vietnamese (vi), Vietnamese – Vietnam (vi-vn)

Sound wise, children's speech also differs greatly from adults' speech in many aspects. One of the first important differences is that children's speech is characterised by the fact that it is generally higher in pitch compared to adult speech (McGowan and Nittrouer, 1988). Other acoustic differences include formant frequency, the average phone duration, the speaking rate, the glottal parameters, and pronunciation (Lee, Potamianos and Narayanan, 1999). Chatbots and language models that have been trained on "normal" corpora will thus often not be able to accurately mimic children. It is of importance to keep this in mind when working with language models or chatbot frameworks for the Talking Child-Avatar.

2.8 The implementation of sentiment analysis

In order to create an accurate chatbot, it is important to understand how the child's emotions change throughout the conversation. There are many chatbots out there that are trained to detect emotions and moods with the help of sentiment analysis (Yin et al., 2019) (Oh et al., 2017) (Kao, Chen and Tsai, 2019) (Zhou et al., 2018). However, in contrast to these chatbots, we do not have annotated sentiment data available. For this project, there are a plenty of transcripts available but these transcripts are not annotated with sentiment labels. Thus, we apply the technique of zero-shot learning. Whenever we talk about zero-shot learning, we mean that we are working with objects that have not been present during training (Xian et al., 2020). More precisely, zero-shot learning recognises new categories of instances without training example because the high level descriptions of the new categories relate them to categories previously learned by the machine, so it learns new classes by just knowing their descriptions (Romera-Paredes and Torr, 2017). In this case, we exploit the zero-shot classifier of the Hugging Face library[5] (Wolf et al., 2020) and a cosine similarity approach with pre-trained models.

Brown et al. (2020) showed that the billion-parameter models available can perform competitively in many domains, data sets, and on different tasks than they are trained on. They need less specific data and are still able to achieve state-of-the-art results, as long as they are trained on a adequately large and diverse dataset. The GPT-2 model outperformed the then-available models on 7 out of 8 tested language modelling datasets. Initially, zero-shot learning (ZSL) was referred to the process of training a model and evaluating it on a completely different task. However, the common practice has shifted slightly, where now the general consensus is that ZSL alludes to training a model and then using it for a task for which it has not been explicitly and solely trained. The difference between these two descriptions is that in the traditional zero-shot learning environment the models require a label for an unknown class.

Zero-shot learning works pretty well, especially in the domain of Natural Language Processing (NLP). The advantage that NLP has over for example the vision domain is that it is fairly easy to model both the class name as well as the to-be-classified input in the same embedding space. The zero-shot pipeline created by Hugging Face use a Natural Language Inference technique to classify premises. The pipeline can utilise the models present in the Hugging Face library but uses *bart-large-mnli* as default. The premise is the sequence that we want to classify, and each candidate label can be seen as a hypothesis. Then both the premise and the hypothesis are put through the model. Each hypothesis needs to run through the model with the premise and, therefore, requires its own forward pass. Since we require only one label as output, the scores for entailment as logits are put through a softmax such that the candidate label scores add to 1. However, if it would be desired that multi class

[5] https://huggingface.co/

classification is true, we can set multi_class=True. Then, instead of using softmax for the scores for whole set of labels, the pipeline will use softmax for each hypothesis individually.

2.9 Similar projects

Ohlheiser and Hao (2021) have developed a chatbot, called Riley, that helps train counsellors at the Trevor Project, a non-profit organisation focused on the prevention of suicide among teens within the LGBTQIA+ community. The term LGBTQIA+ is often used to refer to the whole queer community, as it stands for Lesbian, Gay, Bisexual, Transgender, Queer, Intersex, and Asexual. The objective is to train volunteers more efficiently while simultaneously improving their techniques compared to past training methods. They created a fine-tuned version of the GPT-2 model to create Riley. The chatbot is fine-tuned on data from their own crisis hotline.

Another similar project has the name Robin and is developed by the Technical University of Delft. Robin looks like SpongeBob and can both chat and show emotions. The prototype supports children who are bullied on the Internet. It is able to show sixteen different facial expressions (Bohn Stafleu van Loghum, 2014).

The difference between these project and the Talking Child-Avatar is that they work in a different and, slightly more, informal setting. Robin focuses on improving the general well-being of the children with whom its interact. In contrast, the Talking Child-Avatar is not created to support the children but to train the CPS workers and law enforcement's personnel to interview in a correct way. Our framework, therefore, does not interact directly with the child itself. Where Riley is also created to train personnel, the context is slightly differs as it is aimed at older children and it is used as an emotional support system instead of an interviewing training module.

2.10 Summary

Nowadays, a lot is possible due to the large and extensive language models and chatbot frameworks available. Most language models are promising when it comes to generating human-like language when it comes to adult language in certain contexts. As we do not have enough annotated data available, we cannot train our own RNN, LSTM, or Transformer, and thus, we have to use a pre-trained language model for our language generation. The most promising models are GPT-2 and GPT-3; however, at the beginning of this research project, we did not have access to GPT-3. The language models are advanced and can generate language and their own storylines, but they often need some more annotated data to fine-tune to the setting. The recreation of the setting is extremely important as we try to mimic an abused and neglected child, not just any random adult. Chatbot frameworks can be a solution to the lack of data problem as they are more rigid and easier to set up. However, they require more work to

convert our data to the specific data format since every framework often requires something different. A disadvantage, however, is that they cannot generate their own storyline. The most cost-effective chatbot regarding the functionalities is RASA, as it is open source and thus free to use while it still has many valuable functions. RASA allows us to create storylines within its environment which is a significant advantage.

For the sentiment analysis component of the chatbot, we will be able to use some language models as they have been trained on other data. We can use these language models in a zero-shot setting in order to extract the required emotions. The Hugging Face zero-shot pipeline will be used to do this, among a cosine similarity approach; more about this in Chapter 4.

In the next chapter, we will start by testing out GPT-2 in a chatbot setting. We will compare this with a RASA chatbot. Furthermore, a sentiment pipeline will be created and implemented in the RASA framework.

Chapter 3

Chatbot Framework

In the previous chapter, we saw that GPT-2 is a promising language model for language generation. This chapter tests two versions of GPT-2 as a generative model in a chatbot setting. Furthermore, the development of the Talking Child-Avatar in the RASA environment combined with the creation of a sentiment analysis pipeline is discussed.

3.1 GPT-2

3.1.1 A first look at the possibilities of GPT-2

The most promising language models from Chapter 2 are GPT-2 and GPT-3. At the beginning of this research project, we did not have access to GPT-3, and thus we were forced to run experiments solely with GPT-2. As mentioned before, GPT-2 is a probabilistic-based transformer, that is freely available in the Hugging Face library. We are looking for a language model that can mimic the language characteristics of an abused child and can mimic this in an interview setting. One of the promising aspects of GPT-2 is that it performs exceptionally well on open-ended language generation prompts (Das and Verma, 2020). The GPT-2 model generates language based on the probability distribution of a word sequence. We tried two different GPT-2 models, where the difference lies in the fine-tuning of the model. The two different models are the non-fine-tuned GPT-2 model and DialoGPT (Zhang et al., 2020), a GPT-2 model that has been fine-tuned on Reddit dialogues. Those models are combined with different search strategies. The different search strategies are greedy search, beam search, and a sampling approach. At each time (t), greedy search (Cormen, 2009) simply selects the next word with the highest probability based on all n previous words:

$$word_t = argmax_w P(w|w_{t-n:t-1}) \tag{3.1}$$

A greedy search is prone to miss high-probability sequences as it eliminates all other possibilities. The highest probable option may not be the best one in the future. Beam search reduces this risk by keeping track of an n number of beams. The beams with the highest probability will be kept in memory and updated at every step. When we set the number of beams

to three in the model, we will select three words at each time step and develop them to seek the sequence with the highest overall probability. Not only does it keep track of these other beams, but it is also possible to see the best beams after the generation is complete. This can be useful as we want to choose the beam that best fits our purpose, not necessarily the one with the highest overall probability. Setting the number of beams to one is equivalent to a greedy search. Both greedy search and beam search try to find the sequence with the highest probability. These strategies may work efficiently for tasks where the outcome is predictable, such as machine translation or text summarization. However, it is not always the option for open-ended text generation, where the length of the desired output can vary substantially, as is the case in chatbots. Therefore, we introduce the concept of sampling. Sampling follows a more human-like approach since natural language is also not based solely on probabilities (Holtzman et al., 2020). To put it simply, sampling picks the next word at random but uses a conditional probability distribution to do so, i.e., formally, sampling looks like:

$$w_t \sim P(w|w_{t-n:t-1}) \tag{3.2}$$

In all situations, we must consider the constraint on the number of tokens that the model can process at a time. The largest version of GPT-2, for example, has a fixed length of 1024 tokens, so we cannot calculate

$$word_t = argmax_w P(w|w_{t-n:t-1}) \tag{3.3}$$

directly when t is greater than 1024. Instead, the sentence is divided into subsequences equal to the maximum input size of the model.

We explore the possibilities of the two existing models, GPT-2 and DialoGPT. Both models are implemented and explored in combination with different search strategies. To sum it up:

- The first model is a non-fine-tuned GPT-2 model. We explore this model with greedy search, beam search, and a sampling approach.
- The second model is a GPT-2 model that has been fine-tuned on Reddit dialogues, called DialoGPT. The same search strategies as with the first model are exploited.

3.1.2 An evaluation of GPT-2

Extracts of conversations with GPT-2 and DialoGPT can be found in Figures 3.1, 3.2, 3.3, 3.4, 3.5, and 3.6, respectively. As for the sampling approach, we need to incorporate some randomness into the generating process so that we can sample from a large number of possible sequences based on probability. However, sampling with low probabilities on an extensive collection of sequences can result in random generation.

Based on the example conversations that we had with the non-fine-tuned GPT-2 model, we can conclude that this is not an approach that works. The model does not have the required conversational skills, and

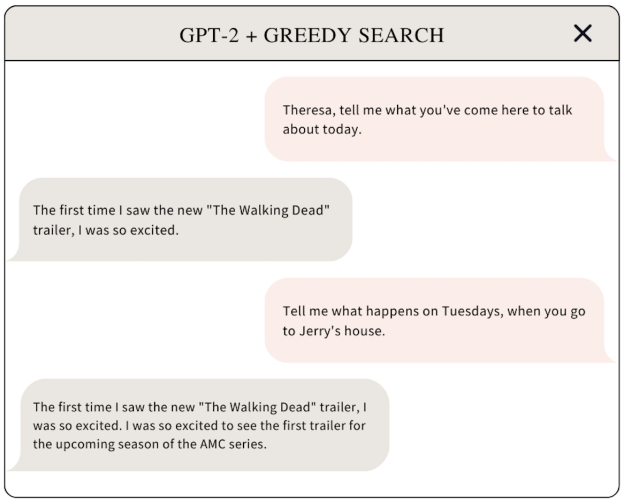

Figure 3.1: A sample conversation with the non-fine-tuned GPT-2 model in combination with greedy search.

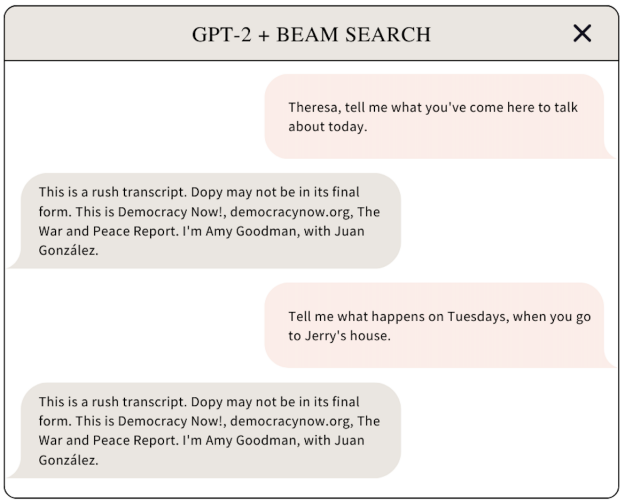

Figure 3.2: A sample conversation with the non-fine-tuned GPT-2 model in combination with beam search.

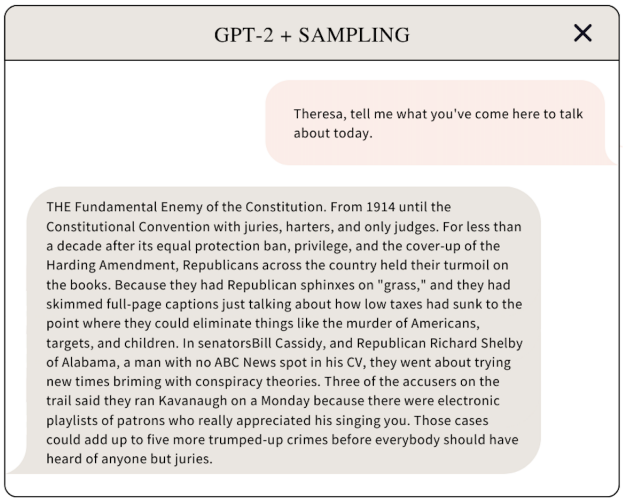

Figure 3.3: A sample conversation with the non-fine-tuned GPT-2 model in combination with sampling.

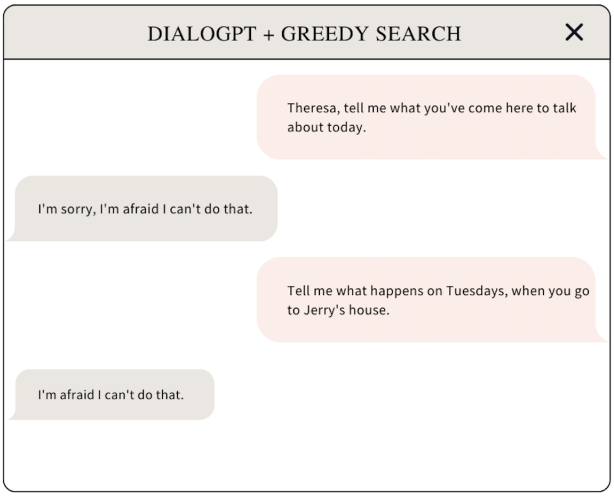

Figure 3.4: A sample conversation with the DialoGPT model in combination with greedy search.

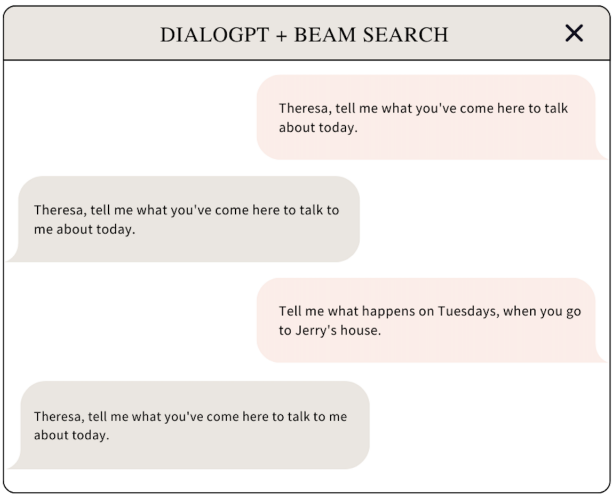

Figure 3.5: A sample conversation with the DialoGPT model in combination with beam search.

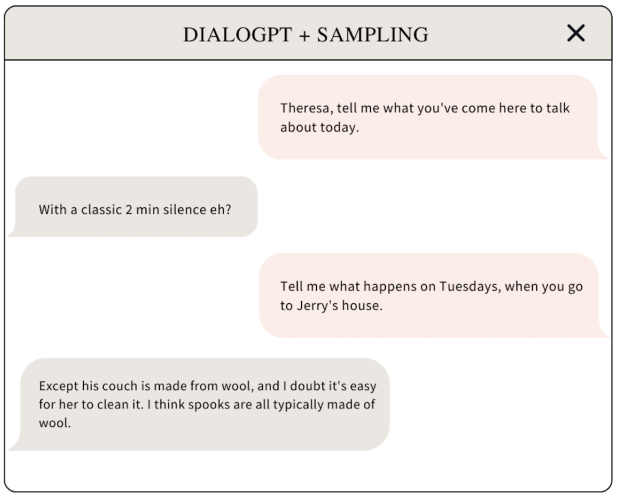

Figure 3.6: A sample conversation with the DialoGPT model in combination with sampling.

the answers do not make sense in context. The text seems grammatically correct, but there are numerous repetitions, and it misses the nuances that we expect from our Talking Child-Avatar. The DialoGPT model performs slightly better, it lacks the refinement of a chatbot trying to mimic a child who has been maltreated.

To further improve the system, it would be possible to implement one of the following approaches:

- Lowering the sampling temperature, which reduces the possibility of selecting low probability words while increasing the likelihood of selecting high probability words.

- Using Top-k sampling instead of selecting all probable occurrences will allow us to exclude low probability terms from being selected.

However, these implementations will improve the output of the chatbot and not necessarily the setting of where the conversation takes place. As mentioned before, mimicking a child is going to be a great challenge since every big model available is trained on adult language, let alone a child that has been traumatised due to abuse and maltreatment. Since we cannot implement our own storylines, we have to move on from GPT-2 to RASA. RASA is not able to generate its own storylines and text, but it will be able to follow the storylines we write and implement. This is a separate component to the sentiment analysis where pre-trained models will be able to accomplish the task.

3.2 RASA

After the first experiments with GPT-2, we explored the possibilities of RASA[1](Hassan et al., 2022). The objective of RASA is to make machine learning based dialogue management and language understanding available for everyone, also for non-machine learning software developers. (Bocklisch et al., 2017). It provides an open-source framework to develop automated solutions for text-based conversations and is built on the TensorFlow (Abadi et al., 2016) framework. In order to develop a dialogue model, RASA utilises various modules that all use different deep learning models. The format of the training data does vary between the different modules. The NLU module jointly predicts the intents and entities. The NLU module utilises the Dual Intent and Entity Transformer (DIET) (Bunk et al., 2020) to do so.

3.2.1 An evaluation of RASA

One of the advantages is that it is relatively easy to implement. Furthermore, RASA makes it possible to create a dialogue system without a great

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

amount of data. The way that this open-source system works is by following a certain script. The programmer has to specify the intents and their corresponding actions. These actions are described in a JSON file. The data needs to be formatted into the required JSON format. It needs to be specified which parts of the data belong to which intents. The problem with this system is that with not enough data, RASA will use the same couple of stories over and over since it is not able to generate new sentences; it can only follow the prescribed intents and storylines. On the other hand, if there is too much data that is somewhat similar, many more intents need to be created. Nevertheless, if there are too many intents, they tend to overlap, which would weaken the intent classification in RASA. Last but not least, a lot of manual effort is required to transform the original data into the format that RASA accepts. Concluding, even with the small amount of data, RASA provides a remarkable environment to control the flow of conversation so the context of an abused child can be insinuated. However, there is a possibility of worsening this environment due to too many storylines and intents.

By use of the *rasa shell* command, it is possible to communicate with the latest trained model through the command-line interface. The flow of conversation from our RASA bot is shown in Figure 3.7. In the diagram, the oval represents the start of the procedure, rectangles represent processes, and the parallelogram symbolises user input. The blue colour stands for the launching of the RASA environment and the user input that gets provided. The yellow colour portrays the *Interpreter*, whereas the red colour represents the *Tracker*. Finally, the green rectangles represent the action, policy, and response decisions. To explain the rest of Figure 3.7, we should start at the blue circle in the top left. The command launches the RASA environment and loads our trained NLU model. The latest trained model is loaded by default, but it is possible to use the *–model* flag and specify a different model that needs to be used instead. Then, the user sends the first message, also called a request. This message is sent to the *Interpreter*, the yellow section of the flowchart. The Interpreter parses the request and extracts the intents and entities. The *Tracker* creates an object of type *UserMessage* from the input, which is of type string. This *UserMessage* is a dict that contains the text, the intent, the intent ranking, and the entities. The intent ranking is the confidence of the intent classification of the classified message. An *UserMessage* may look like this:

```
{
    "text": 'greet{"name":"Rasa"}',
    "intent": {"name": "greet", "confidence": 1.0},
    "intent_ranking": [{"name": "greet", "confidence": 1.0}],
    "entities": [{"entity": "name", "start": 6,
    "end": 21, "value": "Rasa"}],
}
```

After creating this object, the result is sent to the *Tracker*. The *Tracker* keeps track of the state of the conversation. This is an essential part of the RASA environment as conversations are based on storylines and follow a certain scenario. It requires the bot to keep track of where it is

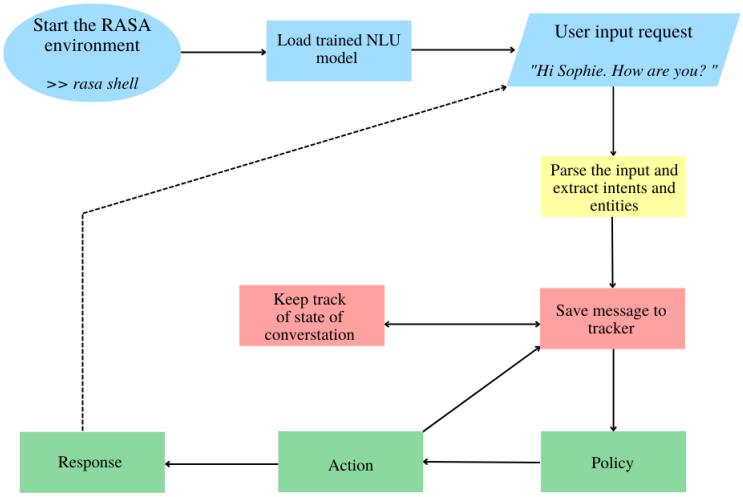

Figure 3.7: The flow of conversation when the command *rasa shell* is run.

in conversation and what the next possible steps can be. The policy gets executed as soon as the *Tracker* has been updated and consolidated. The policy determines what action to take. Lastly, the action gets executed, and the output gets generated. The user receives this response from the system. A user response to this output will trigger the flow of conversation again.

RASA seems like a perfect unparalleled option as there is a lack of sufficient data to develop a solution and train a model from scratch and aim to develop a prototype for proof of concept. We have approximately a thousand transcripts of well-conducted training interviews at the time of writing. Not only is this not enough to train our own deep learning model from scratch, but they are also not annotated, which makes creating a machine learning solution almost impossible. This dataset contains conversations between 5- to 7-year-old children and interviewers. The chatbot is currently designed based on research in interview methodology by the National Institute of Child Health and Human Development's (NICHD) (Martine B. Powell and Brubacher, 2020). It is created for the specific purpose of practising an investigative interview methodology based on those best-practise guidelines. The mock-interview transcripts have some similar storylines; therefore, it is possible to group them and build one persona per cluster.

So concluding, Rasa has all the components in place to create a decent chatbot, but in a very rigid way. The data processing takes up a lot of time, and it is not able to come up with its own storylines. Not only is it not able to create its own storylines, but it is also not capable of answering questions outside of the scope of the predefined story. Consequently, it is able to answer questions about a child's life only if they are written down. However, questions that may not be relevant to the alleged

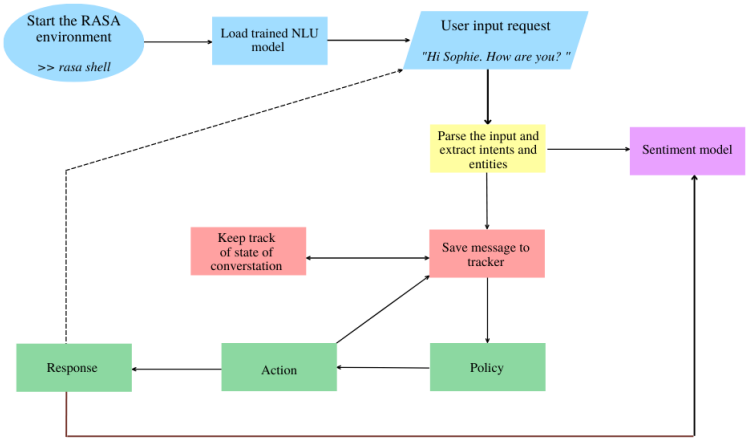

Figure 3.8: The flow of conversation with the integrated emotion pipeline.

incident are left unanswered. It is possible to model small talk and other questions; however, this leads to a large number of intent definitions, which negatively affects the performance of the intent classification model and results in a poorer chatbot. For now, RASA provides a lot of control over how we want the conversation to proceed, but it is not a scalable solution for the future. The aim of the Talking Child-Avatar is to build a chatbot that can show different personas at random, but this is an intermediate solution.

3.2.2 Sentiment pipeline in RASA

The importance of the sentiment component becomes apparent when it is integrated into the complete Talking Child-Avatar. An emotion prediction is only valuable when other parts of the system use the results. The sentiment component plays an integral role in both the visual and audio output by altering the output based on the extracted emotion. An emotion can be used by, for example, changing the facial expressions or the pitch of the voice-based on the specified emotion.

To be able to do so, we created an emotion pipeline that predicts both the emotional valence of both the interviewer's input and the chatbot's output. The flow of conversation can be seen in Figure 3.8. This flow looks very similar to the standard RASA chatbot conversation flow shown in Figure 3.7. The difference between the two figures is that the emotion pipeline flow sends both the users' input and the bot's response to our sentiment model.

There are three ways to interact with the RASA chatbot. We touched upon the first one in Section 3.2.1, namely *rasa shell*. This command opens the RASA environment on the command-line and lets the user communicate with the RASA chatbot. This command also works with the

integrated sentiment pipeline; however, the user will not be able to see the actual classifications even though they do get calculated. Only the input and the bot's responses are visible. The second command to interact with the RASA chatbot is *rasa shell nlu*. Running this command will provide the user access to the RASA NLU interface. Using this interface, the user will be able to see all intent classifications of the model, including the sentiment entity of the user's input. Due to the closed-off RASA environment, it is not possible to directly classify the bot's responses in this environment. However, this is possible by running the following three commands in three different terminal windows:

- rasa run -m models –endpoints endpoints.yml –port 5002 –credentials credentials.yml
- rasa run actions
- python3 start.py

The combinations of these three commands set up a RASA server at a localhost address, run the actions file, and run a script that connects to the localhost. Using the Python requests package[2] to connect to the localhost address where the RASA bot is running, we are able to receive and classify both the human input as well as the RASA output. This is necessary because classifying the RASA response is of great importance to our system, as this part mimics the child. The emotion pipeline receives these texts as input and then provides the input for the audio and visual parts. A small conversation between a user and the RASA chatbot looks similar to this, where the users' input and the bot's response alternate.

```
1 Hello Sophie. How are you today?
2 message sentiment: {'sequence': 'Hello Sophie. How are you today? ',
      'labels': ['enjoyment', 'anger', 'sadness', 'fear'], 'scores':
      [0.6531327962875366, 0.11979195475578308, 0.11764132231473923,
      0.109433911740779881}
3 I'm fine.
4 bot response sentiment: {'sequence': "I'm fine.", 'labels': ['enjoyment',
      'anger', 'fear', 'sadness'], 'scores': [0.6787578463554382,
      0.113361656665802, 0.1127132847905159, 0.09516724199056625]}
5 Do you know why you are here today?
6 message sentiment: {'sequence': 'Do you know why you are here today? ',
      'labels': ['sadness', 'anger', 'enjoyment', 'fear'], 'scores':
      [0.5257893204689026, 0.21673724055290222, 0.13353712856769562,
      0.123936355113983151}
7 About my teacher.
8 bot response sentiment: {'sequence': 'About my teacher.', 'labels':
      ['sadness', 'anger', 'enjoyment', 'fear'], 'scores':
      [0.3919742703437805, 0.2765852212905884, 0.19014804065227509,
      0.1412924975156784]}
```

The labels represent the specified emotions, whereas the scores represent the probabilities of said emotions, respectively. The final model that will be used for the sentiment classification is not set in stone yet and will be elaborated on in the following chapters.

[2] https://github.com/psf/requests

3.3 GPT-3

As mentioned above, we aim to create a chatbot with different personas. A way to achieve this would be GPT-3. With more annotated transcripts available, it will be possible to fine-tune GPT-3, which will allow us to capture the storylines of different children and, based on this, design different personas dynamically. The questions in the transcripts will be annotated within a scheme of 15 different categories. The children's responses, on the other hand, will be annotated as either productive or non-productive. This data will be used to develop a deterministic model that can be used in order to create a feedback mechanism. This feedback mechanism will, in time, be able to regulate GPT-3 in real-time to alter the behaviour of the child's responses based on the type of question that the interviewer asks. GPT-3 is also capable of generating language, just as its predecessor, GPT-2, which would make the model less rigid and more random. This random aspect is excellent when it comes to a virtual training avatar, as the trainees will be exposed to new situations every time. Therefore, we believe that GPT-3 can solve the issues that arise with RASA, as we expect that it can converse in a coherent way while also generating its own story.

3.4 Summary

GPT-2 is not sufficient as a conversation architecture. It is not able to grip the complicated interview setting. The results may improve when annotated data is available to fine-tune the model, but for now, language models do not seem adequate enough to use. Therefore, we moved on to a chatbot framework where it is easier to steer the conversation in a particular direction and simulate the setting that we want. RASA is a perfect solution for now but lacks the deep language understanding that we desire the Talking Child-Avatar to have. It is not able to generate its own storylines like a language model would be able to, and it gets messy when implementing small talk. GPT-3 may be a solution in the future.

The sentiment pipeline in RASA is very promising. It is capable of calculating both the emotion of the interviewer and the interviewee. The model that is going to do the classification in the end will be decided later; more about this will be discussed in Chapter 6 and 8.

Chapter 4

Sentiment

This chapter explores the models and strategies from Chapter 2 for sentiment extraction in more detail.

As previously mentioned, most projects have access to vast amounts of annotated data to train and test their systems. Unfortunately, we do not have that luxury, nor are similar open-source datasets available. However, we do have access to a thousand mock interview transcripts provided by the Centre for Investigative Interviewing at Griffith University, Australia (Martine B Powell, Guadagno and Benson, 2016). These mock interviews are created as part of their investigative interview training for mainly social workers, police, and psychologists. In these mock interviews, a trained actor mimics an allegedly abused child. Real-life child investigative interviews will be added to the system at a later stage in the process. Collecting authentic transcribed investigative interviews with alleged victims of abuse and abuse requires the project to comply with special ethical requirements for sensitive data with vulnerable children. CPS workers or law enforcement personnel will anonymise these transcripts by removing all direct and indirect personal identification information. These mock interviews are not annotated at all; therefore, we use the pre-trained models in a zero-shot learning environment. Initially, the plan was to start with a classification of either neutral, negative, or positive, based on one sentence at a time. We quickly discovered that this would not function accordingly due to the fact that it is difficult to classify a single sentence in one of these three categories. We then moved on to a one-sentence-at-a-time classification using the seven basic emotions: *enjoyment, surprise, fear, sadness, anger, disgust, and contempt*. After the initial experiments, we switched from these emotions to a subset of these, namely the four core emotions: *enjoyment, fear, sadness, and anger*.

The set of emotions is kept small to make it easier to predict the class to which the emotion belongs. Adding even more sentiment classes will make the system more unreliable, as it can be challenging to differentiate between similar emotions. This would result in a pipeline that is unpredictable in classifying emotions. To make more accurate predictions, we also experimented with whole-story-so-far classification. We started by trying

to classify one sentence based on the whole story so far combined with the not yet classified sentence. Not only did we classify sentences based on the whole story or based on that single sentence, but we also experimented with a compromise between the two, namely a sliding window classification. The sliding window classification entails that we only keep a couple of sentences as context. The experiments started with different window sizes of 5, 10, and 15 sentences. Whenever the sentiment of a new sentence needs to be predicted, we only take into account as many sentences as the window size. It was immediately apparent that these results were promising and actually made more sense than the results of the previous experiments. Windows sizes of 5 and 10 yielded better results than a window size 15, and therefore, we also added experiments with a window of size 7.

With these settings, the model was reasonably accurate in predicting the sentiment of a single sentence based on the sliding window. Some sentences are an obvious turning point in the emotional storytelling of a child. These sentences had a high probability for one of the classes,but it was not enough to turn the classification around due to the other sentences within the sliding window. Therefore, we implemented a threshold that disregarded the context if, and only if, the probability of one specific class exceeded this threshold. The window for the following sentence will only take into account this specific sentence so that the change is noticed from then on. Based on the observed probabilities, there have been experiments with thresholds of 0.4, 0.5, and 0.6.

4.1 Approaches

We will experiment with a sequence embedding model that calculates the cosine similarity, as well as the Hugging Face zero-shot pipeline that performs classification based on Natural Language Inference.

4.1.1 Cosine similarity

The Cosine similarity measures the similarity between a sentence and a class by measuring the similarity of the respective vectors based on the cosine of the angle between them (H.Gomaa and A. Fahmy, 2013) (Lahitani, Permanasari and Setiawan, 2016).

Every sentence, also called premise, gets classified based on the highest cosine similarity compared to every class. The formal definition looks like this:

$$\hat{c} = argmaxcos_{c \in C}(\Phi_{sent}(x), \Phi_{sent}(c)) \tag{4.1}$$

$\hat{c}$ is one possible class out of all classes C. We calculate the cosine similarity, cos, based on the sentence and the class. The tokenizer will batch encode the list of premises and pad them to the maximum length. After the padding, we run them through the model and get the tensors back. These tensors will be mean-pooled over, so we get the sequence-level representations. Mean pooling implies the calculation of getting the

average for the whole set of sentences. The result of the cosine similarity approach will be a tensor with a cosine score for each emotion; the higher the number, the more similar the emotion is to the input sequence.

4.1.2 Hugging Face zero-shot pipeline

The Hugging Face zero-shot pipeline is based on Natural Language Inference (NLI). NLI not only calculates similarity, but it can also report on the degree of compatibility of two sequences. It works with both a premise and a hypothesis instead of just premises and class labels. As this is the case, each premise and each class label requires its own forward pass through the model. The pipeline has an option to use multi-class classification or not, meaning that there are multiple true labels or only one. We chose to set multiple true labels to false for our experiments; this is also the default option. Using this setting, and due to the softmax function, all scores for the labels will add up to 1, as the softmax turns logits into probabilities. The softmax function is formally written as:

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{4.2}$$

If multi-class labels were set to True, each candidate label would be individually scored by the softmax function. Since that is not the case, the result will be a list of probabilities corresponding to the different emotions. The higher the number, the more sure the model is about its classification.

4.2 Models

Using both the cosine similarity and the NLI approach, we use different pre-trained models to test them. We described some state-of-the-art models in Section 2.4, but not all of these models are freely available. Based on what models are freely accessible and their noted scores on various downstream tasks, we decided to run the experiments with DeBERTa (Section 2.4.12), DistilBERT (Section 2.4.13), BART (Section 2.4.14), and GPT-3 (Section 2.4.7).

We already explained GPT-3 in Section 2.4. But where we did not have access to GPT-3 at the beginning of this project, we received an API key later on. We got access from OpenAI to exploit the model in their environment. It is, therefore, not possible to test it in either a cosine similarity or a zero-shot setting, but we merely get the results straight from their API. We connect to the *text-davinci-001* engine. OpenAI is the company that developed GPT-3, among other models, and they created engines to describe and connect the models. There are for different engines to connect to the GPT-3 models; these are *Davinci, Curie, Babbage, and Ada*[1]. The most powerful and capable engine is *Davinci*; it is able to execute every task that the other models can execute as well, but often with even less

[1] https://beta.openai.com/docs/engines/gpt-3

instruction. It is the prime engine when it comes to jobs that require a lot of understanding of the content, like summarising, solving logic problems, or generating creative content. *Davinci* is the most expensive model when it comes to API calls due to the use of more computational resources. *Curie* is another compelling engine, and it is also very fast. *Davinci* and *Curie* are known as the two most powerful engines, capable of understanding texts on a deeper level. The latter two engines, *Babbage* and *Ada*, are faster but do not have the competence compared to *Davinci* and *Curie*. *Babbage* and *Ada* are very useful when it comes to performing straightforward tasks. However, for our purpose, we require more textual understanding proficiency from the engine. We use *Davinci* as it is the capable GPT-3 model, and OpenAI recommends to start experimenting with this model in the beginning.

Chapter 5

User Studies

By conducting user studies we want to understand how humans will annotate certain interview excerpts from the interviews. The three different user studies all start with an explanation of the research objective, followed by a trigger warning since the questions can be hard to handle. After this introduction, there are three personal questions. We ask the age, their level of education, and their gender. The aim is to form a diverse group of participants of different genders, age groups, and levels of education, since we want it to be representative of the population as a whole. These questions are followed by a description of the different emotions. The description that has been given in to the participants in the first user study is based on the explanations given on Paul Ekmans website[1]. The emotions were described as following:

- *Enjoyment* is, for many, the most desirable of the seven universal emotions, which usually arises from connection or sensory pleasure. It describes a family of pleasurable states, everything from peace to ecstasy.
- *Surprise* arises when we encounter sudden and unexpected sounds or movements. Its function is to focus our attention on determining what is happening and whether or not it is dangerous.
- *Fear* arises with the threat of harm, either physical, emotional, or psychological, real or imagined. While traditionally considered a "negative" emotion, fear actually serves an important role in keeping us safe as it mobilises us to cope with potential danger.
- *Sadness* results from the loss of someone or something important. What causes us sadness varies greatly based on personal and cultural notions of loss. Sadness describes the range of emotional states we can experience containing everything from mild disappointment to extreme despair and anguish.
- *Anger* arises when we are blocked from pursuing a goal and/or treated unfairly. communicates anything from mere dissatisfaction to threats.

[1] https://www.paulekman.com/universal-emotions/

- *Disgust* arises as a feeling of aversion towards something offensive. It contains a range of states with varying intensities ranging from mild dislike to intense loathing. All states of disgust are triggered by the feeling that something is averse, repulsive and/or toxic.
- *Contempt* is the feeling of dislike for and superiority (usually morally) over another person, group of people, and/or their actions.

It is important to note that contempt is related to but different from disgust. Although both contempt and disgust can be directed toward people and their actions, disgust can also be aroused by objects that are aversive to the senses (taste, smell, sight, sound, and touch). Additionally, contempt includes the feeling of superiority over the target of contempt, whereas one does not necessarily feel superior to the person/thing that disgusts them.

The first user study uses 7 different emotions, while the second and third user studies only use 4. The sets of emotions follow the available research, as explained in Section 2.2. The above mentioned description of the emotions is found in the first user study, as that is the only one that presents the seven universal emotions, that are *enjoyment, sadness, anger, disgust, contempt, fear, and surprise*. In contrast, the second and third user study only consist of the emotions *enjoyment, sadness, anger, and fear*. The descriptions in both of these user studies are changed accordingly. All user studies conducted consist of only closed-ended questions, intended to conduct quantitative research. These closed-ended questions can be transformed into numerical data that we will statistically analyse later to find patterns, trends, and correlations within our data. The user study is distributed among native and non-native English speakers. Then, single-sentence excerpts are presented. The user sees a sentence and has to classify it as one of the emotions that have been described in the previous section. The first and second studies have the same 13 sentences, while the third has 8 different ones. The single sentence excerpts give us a good insight into whether the model can adequately predict sentiment when no context has been taken into account. Thereafter, the human gets the same kind of questions, but instead of having to classify a single sentence, they have to classify a conversation of a certain window size. In the first and second studies, we only display excerpts of window size 5. Both user studies portray the same 9 excerpts. Therefore, the only difference between the first and second studies is the number of emotions that the user can choose between. On the contrary, the third study has 36 different window excerpts. There are 12 excerpts shown with window size 3, 12 with window size 5, and 12 with window size 7.

As mentioned in Section 1.5, there are a few limitations concerning the conducted user studies. The first limitation of this study is the insufficient sample size of the user studies conducted. Furthermore, this study will not encompass extracting all possible emotions, but will focus solely on classifying a predefined subset. Another limitation of studies is that it can be difficult for humans to extract emotional responses from a single sentence or excerpt, which may be due to the way the questions are

organised since it was not possible to choose multiple emotions or 'none of the above'. This choice was made because the model always outputs the most probable emotion as well.

In the next three sections, we will discuss the three different user studies in more detail. In addition, we review the results and discuss them. The full list of questions present in all three user studies can be found in Appendix A. A complete overview of the graphs showing the distributions of the responses given to the different user studies can be found in Appendix B.

5.1 User Study 1

As described in the previous section, the first study consisted of 22 questions, 13 of them are single sentence excerpts, and the remaining 9 questions are excerpts of window size 5. The participants had to choose between seven different emotions. These emotions are *enjoyment, surprise fear, sadness, anger, disgust, and contempt*. This user study was conducted as a preliminary user study to see if seven emotions would be sufficient or if it would make the choice overly complicated. There were twelve participants. 41.7% of the participants was between the age of 18 and 25, 50% between 19 and 35, and 8.3% was between 56 and 65. The 91.7% majority was female compared to 8.3% male. All participants graduated at least high school (16.7%), and the highest finished level of education was a Master's degree (25%). The remaining 58.3% acquired a Bachelor's degree as highest educational qualification.

In the first user study, there was unanimous agreement in only one of the twenty-one questions, that one being a single sentence excerpt. This answer distribution for this question can be seen in Figure 5.1.

As said before, this was the only question where there was a collective agreement, most answer distributions look like the ones in Figure 5.2, Figure 5.3, and Figure 5.4.

We mark the emotion as a clear winner, if there was a difference of 20% or more between the winning class and the second place. 9 out of 13 single-sentence excerpts have a clear winner, compared to 5 out of 9 story excerpts with window size 5.

A critical measurement in qualitative research is integral to research dealing with subjectivity and, therefore, reliability. If we present our subjects with a questionnaire and ask them to classify the excerpts as a certain emotion, we get subjective ratings. It is important to know to what degree of confidence we can assign these ratings. In statistics, the Intraclass Correlation Coefficient (ICC) is a descriptive statistic that can be used when quantitative measurements are made on units that are organised into groups. It depicts how strongly units in the same group resemble each other. The ICC describes how strongly elements in the same group resemble each other. An ICC between 0.40 and 0.59 is a moderate score. However, it is preferable to have an ICC score of at least 0.75 (Koo and Li, 2016). Despite the fact that there was a clear winner in fourteen questions, the ICC

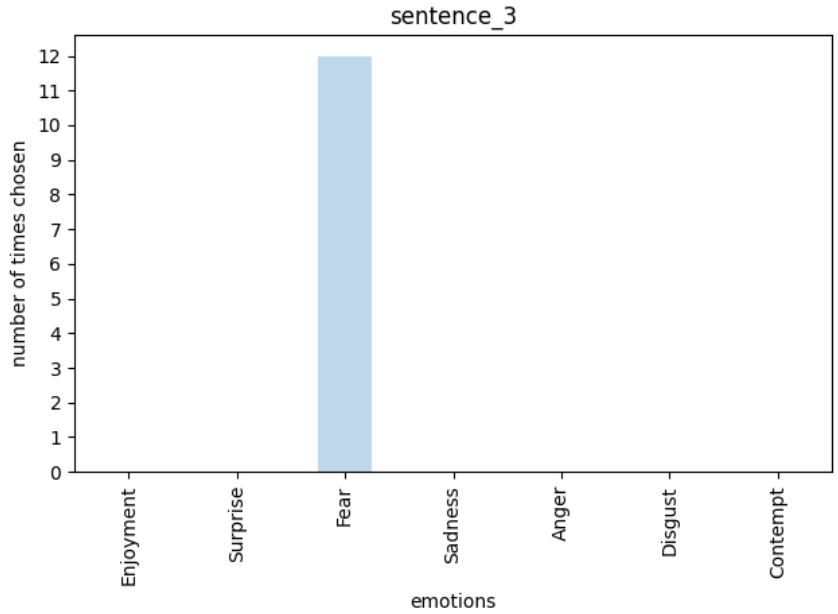

Figure 5.1: The answer distribution for the sentence *"Because I was scared I would get stuck inside."* in user study 1.

based on average measures was low in 7 dimensions, scoring only 0.537 on average measures. The first user study showed us that it is challenging for humans to reach a consensus when classifying data excerpts and having to choose between seven emotions. The involvement of context does not make it easier either.

For a complete overview of the answers per question, Appendix B can be consulted.

5.2 User Study 2

The second study was conducted as a pre-study to the third study, and therefore, only had five respondents. 60% of the participants was female, compared to 40% male. They were either between 18 and 25 years old (60%) or between 26 and 35 years old (40%). 20% had a high school diploma as the highest degree received, for 20% this was a bachelor's degree, and 60% had a master's as the highest level of school completed. Since it was carried out as a preliminary study, there was no need for any more participants, since the objective was to see if the agreement between the participants would be greater if we reduced the options from seven emotions to four.

The first and the second user study consisted of the same questions; the only difference is the number of emotions that the participants had to choose between. In the second user study, we reduced the number of options from seven to four. The sole purpose of this study was to see whether the agreement would improve by changing this. After only

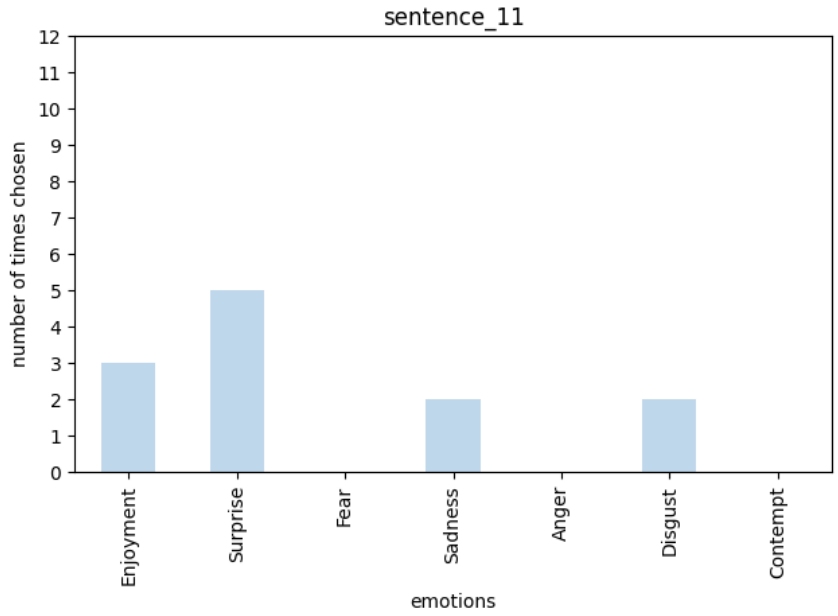

Figure 5.2: The answer distribution for the sentence *"So I touched it."* in user study 1.

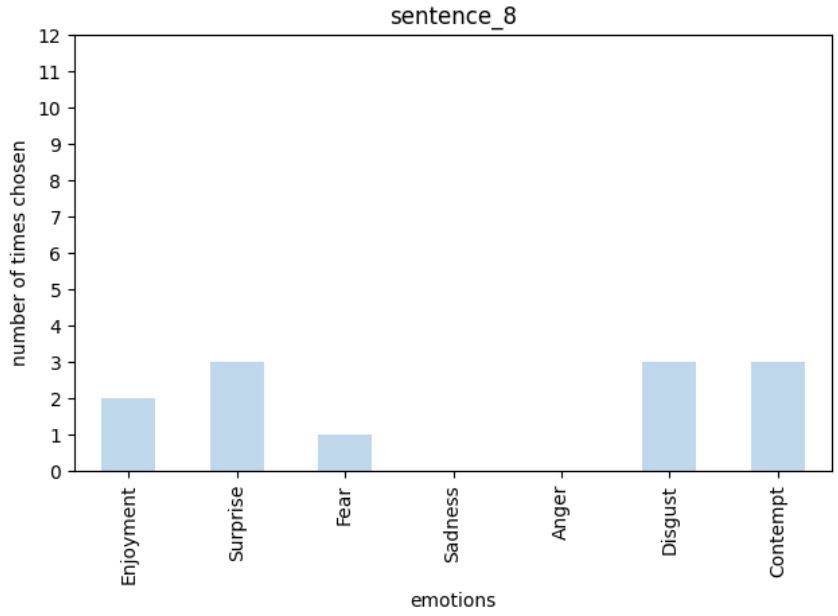

Figure 5.3: The answer distribution for the sentence *"Just in the art room."* in user study 1.

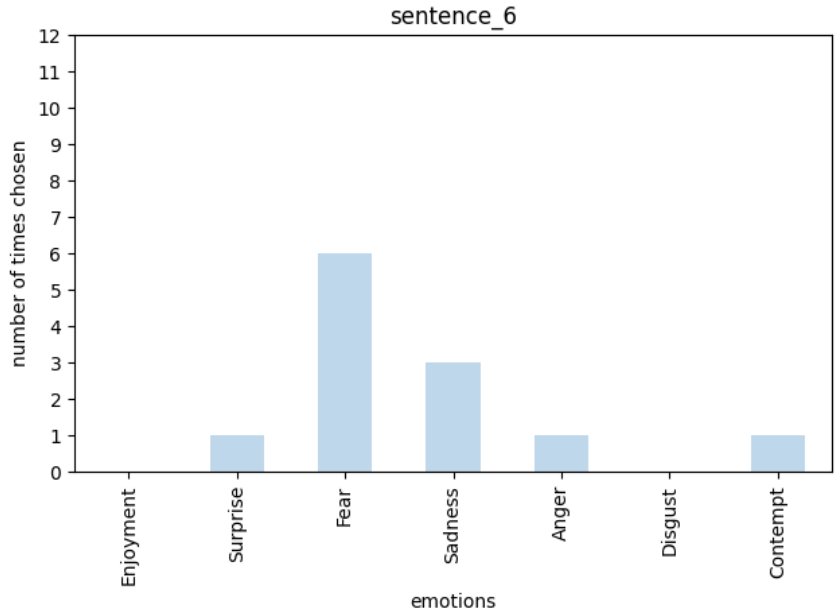

Figure 5.4: The answer distribution for the sentence *"And then he ran away and then an ambulance came."* in user study 1.

receiving five responses, it became clear to us that it indeed improves the results as the distributions were less deviant. The number of clear winners improved from 9 to 10 for the single-sentence excerpts and from 5 to 7 for the story excerpts with window size 5.

These findings were used later in the third study that was more comprehensive than the first two.

5.3 User Study 3

The last user study is the most complete one. It was the most comprehensive, as it contained single-sentence excerpts, as well as excerpts with window sizes 3, 5, and 7. There were 21 participants in total. 52.4% of the participants identified as female and 47.6% as male. Most of the participants were between the age of 26 and 35, but other age groups, such as people between 18 and 25, 36 and 45, between 56 and 65 years old, also participated. The level of schooling differed greatly from no schooling completed (4.8%) and high school as the highest level (42.9%) to a bachelor's (33.3%) and master's (19%) degree.

Where the first and the second user study were conducted as preliminary studies, the third one was set out to be a comprehensive study of how humans would annotate different excerpts from our data. Following the findings from the second user study, participants had to choose between only four emotions. The number of clear winners are 5 out of 8 for single sentence excerpts, and 26 out of 36 for the story excerpts. There were 8 clear winners for the excerpts with window size 3 and 5, and 10 clear winners

for the window size 7 excerpts.

The ICC in four dimensions increased from 0.537 to 0.788 compared to the same measure in seven dimensions.

The different window sizes do have an impact on human judgement, as the ICC differs greatly between them. In the single sentence examples, the ICC is 0.218 with a lower bound of the 95% confidence interval at -0.089 and the upper bound at 0.545. The 95% confidence interval is the range of values on which you can be confident that 95% contain the accurate population mean. The significance is 0.075. The significance is the relationship between the variables. If the number is less than 0.05, the relationship is not statistically signification. So, for the single sentence classification, we can state that it is statistically significant, but receives a fairly low ICC score. This is interesting and explainable due to the fact that there are a lot of clear winners but the disagreement is strong among other classes. The ICC increases to 0.467 in the window size 3 setting, and increases even more to 0.659 with the window size 5 excerpts. It decreases a bit to 0.556 with the window size of 7. We do see that the more information available, the easier it is to judge. However, if there is too much context, it becomes vague again and thus the quality of the judgement declines. It is interesting to see that the window size of great influence to the reliability and unity of human extraction of emotions.

Merging the different categories in *enjoyment* versus the rest would be an interesting observation in terms of ICC scores. The idea behind it would be that *fear, sadness, and anger* are similar categories since they reflect the unhappy emotions compared to the happy emotion of enjoyment. However, the ICC decreased drastically. An explanation for this would be that it is not as easy as mapping the emotions to either happy or unhappy. Only a new study can explore the scores in a two dimension setting.

Chapter 6

Results and evaluation

6.1 Initial results

None of the user studies included excerpts that represented the entire conversation between an interviewer and a child. This choice was made due to the poor results in the early stages. However, it is essential to provide an explanation as to why the whole story excerpts were scraped from the studies. Each row in Figure 6.1 represents a single sentence that is classified. As can be seen, the single-sentence classifications differ significantly from one to the next one. In contrast, the whole-story classification has too little differentiation throughout the classification, as it gets stuck on a certain emotion. Whereas the single-sentence classification may miss the significance of context, the opposite occurs when predicting the emotion based on the whole-story-so-far. The classification of this sentence is being influenced too much by the other emotions that have been present before. Thus the model is not able to pick up subtle emotional changes. By way of illustration, if more than twenty sad sentences are already present in the story, and the sentiment starts to switch to surprise, these twenty sentences will have a more significant weight than that one surprise sentence. It is almost impossible to outweigh that many sentences with just one new sentence. Furthermore, there is a problem of having too many sentences with too many different emotions. If the story is relatively long, all types of emotions can be present within the story. If this is almost equally divided, we are almost doing the same classification as one-sentence-at-a-time classification, meaning context is not being utilised.

Single-sentence classification	Whole-story classification
sadness	sadness
surprise	surprise
relief	surprise
surprise	relief
joy	surprise
joy	surprise
anxiety	surprise
joy	joy
surprise	surprise
surprise	surprise
relief	surprise
surprise	surprise
disgust	disgust
relief	disgust
surprise	disgust
sadness	sadness
relief	sadness
sadness	sadness
disgust	sadness
anger	sadness
sadness	sadness
joy	sadness
relief	sadness
surprise	sadness
surprise	sadness
surprise	sadness
relief	relief

Single-sentence classification	Whole-story classification
anger	sadness
disgust	sadness
relief	disgust
relief	disgust
surprise	disgust
relief	disgust
sadness	disgust
disgust	disgust
sadness	disgust
sadness	disgust
surprise	disgust
relief	disgust
surprise	disgust
relief	disgust
relief	disgust
relief	disgust
joy	disgust
relief	disgust
joy	disgust
relief	disgust
anxiety	disgust
surprise	disgust
surprise	disgust
relief	disgust
surprise	disgust
joy	disgust

Table 6.1: The difference between single-sentence classification and whole-story-so-far classification.

6.2 Comparison user study 3 with Hugging Face models

In this section, the most comprehensive user study, namely number 3, will be compared with GPT-3, the different Hugging Face models combined with the two different strategies to use them. Each header in every table is named either sent_*number* or story_*number*. These numbers correspond to the question in user study 3, where they represent sentences and stories respectively.

6.2.1 Single-sentence excerpts

	Sent_1	Sent_2	Sent_3	Sent_4	Sent_5	Sent_6	Sent_7	Sent_8	Percentage correct
User study	Fear	Fear	Fear	Sadness	Enjoyment	Fear	Fear	Sadness	
GPT-3	Fear	Anger	Anger	Anger	Enjoyment	Fear	Fear	Fear	50%
Cosine Similarity									
bart-large-mnli	Sadness	Sadness	Sadness	Sadness	Anger	Anger	Sadness	Sadness	25%
distilbert-base -uncased-mnli	Fear	Enjoyment	Fear	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	37,5%
deberta-large -mnli-zero-cls	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Sadness	25%
Percentage correct (cos sim)	33,3%	0%	33,3%	100%	0%	0%	0%	66,7%	
Zero-shot pipeline									
bart-large-mnli	Anger	Fear	Anger	Sadness	Enjoyment	Fear	Fear	Sadness	75%
distilbert-base -uncased-mnli	Sadness	Enjoyment	Anger	Sadness	Enjoyment	Sadness	Fear	Sadness	50%
deberta-large -mnli-zero-cls	Anger	Anger	Anger	Sadness	Enjoyment	Fear	Fear	Sadness	62,5%
Percentage correct (zsl)	0%	33,3%	0%	100%	100%	66,7%	100%	100%	
Percentage correct (gpt3 + cos sim + zsl)	28,6%	14,3%	14,3%	85,7%	57,1%	42,9%	57,1%	71,4%	

Table 6.2: Comparison of the results on single sentence excerpts.

The one-sentence-at-a-time classification does work well compared to human agreement, as can be seen in Table 6.2. The models score relatively high when their outputs are compared to the human annotations. The scores reported in this table compared to the the scores from the window size tables, i.e., Table 6.6, Table 6.7, and Table 6.8 are considerably higher. The *BART large* model in the zero-shot pipeline scores the best out of all the strategies we tried. We can also see that the zero-shot pipeline approach yields better results than the cosine similarity strategy. None of the models scored a 100% compared to human annotation, but this is not uncanny as humans were also not able to reach an unanimous consensus for all of the questions. One of the sentences in the user study was *"We watched a movie and then we got some ice cream and then we went to bed."*. The human consensus, the GPT-3 model, and all models using the zero-shot pipeline considered this as *enjoyment*. It becomes more difficult to classify sentences such as *"It really hurt"*, which the zero-shot BART model and the human participants classified as *sadness*, while GPT-3 considers it to be *anger*.

Even though the results are promising, this classification does not meet our needs for a more complicated setting. The context is often crucial in deciding the sentiment of a sentence and thus it is not valuable to classify a single sentence. An example can be an answer that exists solely of the word *"Yeah"*. Without any context, this word is probably joyous. However, if the previous dialogue has been about some kind of maltreatment the child has endured, then it would be inappropriate to classify this as *enjoyment*. Another example is the sentence *"I was on the playground"*. This sentence is innocent and suggests enjoyment when viewed on its own; but in the context of abuse it may no longer have that positive connotation. So even if the models scored really well compared to the human opinion, the results are not useful in the bigger context as it misses the importance of the setting wherein the conversation takes place.

6.2.2 Window size excerpts without questions from the interviewer

The threshold results were not promising, in contrast to the initial hypothesis. The hypothesis, as formulated in Chapter 4, was that a threshold would be able to detect changes in emotion by disregarding the context if and only if the sentence had a sentiment prediction that exceeded the determined threshold. We conducted experiments with threshold of 0.4, 0.5, and 0.6, in combination with the zero-shot pipeline from the Huggingface library. The pipeline ran every classification through the softmax function, what leaves us with probabilities, so a score of 0.55 means that the model predicts a 55% chance of this specific emotion for that specific sequence. In reality, this worsened the emotional classification in some cases as can be observed by comparing the results in Table 6.3, Table 6.4, and Table 6.5 to the results in Table 6.6, Table 6.7, and Table 6.8. However, in most cases it did not affect the scores whatsoever. Even if the scores improved, as happened with two models in combination with a window size 3 and a threshold of 0.4, it did not increase significantly. The threshold strategy did not seem to work as well as we expected due to many miss-classifications for the *enjoyment* emotion. Lots of sentences have a positive connotation when seen on their own, but not in the broader picture, as is also the case with the sentence *"I was on the playground"* discussed in the previous section. The single-sentence classification deemed too influential by using a classification threshold.

	Story_1	**Story_2**	**Story_3**	**Story_4**	**Story_5**	**Story_6**	**Story_7**	**Story_8**	**Story_9**	**Story_10**	**Story_11**	**Story_12**	Percentage correct
User study	Fear	Fear	Fear	Fear	Fear	Sadness	Sadness	Enjoyment	Fear	Fear	Fear	Fear	
Zero-shot pipeline													
bart-large-mnli	Anger	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Sadness	Sadness	Anger	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Sadness	Sadness	Enjoyment	25%
deberta-large -mnli-zero-cls	Sadness	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Ejoyment	Sadness	Fear	Fear	33,3%
Percentage correct (zsl)	0%	0%	0%	0%	0%	100%	33,3%	100%	0%	33,3%	66,6%	66,6%	
Zero-shot pipeline including 0.4 threshold													
bart-large-mnli	Anger	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Sadness	Sadness	Anger	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Sadness	Sadness	Fear	33,3%
deberta-large -mnli-zero-cls	Sadness	Anger	Anger	Enjoyment	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Sadness	Fear	Fear	41,7%
Percentage correct (zsl)	0%	0%	0%	0%	0%	100%	66,7%	100%	0%	33,3%	66,7%	100%	
Zero-shot pipeline including 0.5 threshold													
bart-large-mnli	Anger	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Sadness	Sadness	Anger	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Sadness	Sadness	Fear	33,3%
deberta-large -mnli-zero-cls	Sadness	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Fear	Fear	33,3%
Percentage correct (zsl)	0%	0%	0%	0%	0%	100%	33,3%	100%	0%	33,3%	66,7%	100%	
Zero-shot pipeline including 0.6 threshold													
bart-large-mnli	Anger	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Sadness	Sadness	Anger	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Sadness	Sadness	Fear	33,3%
deberta-large -mnli-zero-cls	Sadness	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Fear	Fear	33,3%
Percentage correct (zsl)	0%	0%	0%	0%	0%	100%	33,3%	100%	0%	33,3%	66,7%	100%	

Table 6.3: Comparison of the results on excerpts of window size 3 with thresholds.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Sadness	Sadness	Fear	Fear	Sadness	Sadness	Sadness	Fear	Sadness	Fear	Fear	Fear	
Zero-shot pipeline													
bart-large-mnli	Enjoyment	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Enjoyment	41,7%
deberta-large -mnli-zero-cls	Enjoyment	Anger	Anger	Fear	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Fear	Fear	41,7%
Percentage correct (zsl)	33,3%	33,3%	0%	33,3%	100%	100%	33,3%	0%	0%	33,3%	66,7%	66,7%	
Zero-shot pipeline including 0.4 threshold													
bart-large-mnli	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Enjoyment	Sadness	Enjoyment	Sadness	Enjoyment	Sadness	Enjoyment	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	25%
deberta-large -mnli-zero-cls	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Sadness	Fear	Fear	25%
Percentage correct (zsl + 0.4 threshold)	0%	33,3%	0%	66,7%	0%	33,3%	0%	33,3%	0%	33,3%	66,7%	100%	
Zero-shot pipeline including 0.5 threshold													
bart-large-mnli	Enjoyment	Anger	Enjoyment	Fear	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	33,3%
distilbert-base -uncased-mnli	Enjoyment	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	41,7%
deberta-large -mnli-zero-cls	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Sadness	Fear	Fear	25%
Percentage correct (zsl + 0.5 threshold)	0%	33,3%	0%	66,7%	33,3%	33,3%	33,3%	0%	0%	33,3%	66,7%	100%	
Zero-shot pipeline including 0.6 threshold													
bart-large-mnli	Enjoyment	Anger	Enjoyment	Enjoyment	Sadness	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	33,3%
distilbert-base -uncased-mnli	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	50%
deberta-large -mnli-zero-cls	Enjoyment	Anger	Enjoyment	Fear	Sadness	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Sadness	Fear	Fear	33,3%
Percentage correct (zsl + 0.6 threshold)	33,3%	33,3%	0%	33,3%	100%	33,3%	33,3%	0%	0%	33,3%	66,7%	100%	

Table 6.4: Comparison of the results on excerpts of window size 5 with thresholds.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Fear	Fear	Fear	Fear	Sadness	Sadness	Fear	Enjoyment	Enjoyment	Fear	Sadness	Fear	
Zero-shot pipeline													
bart-large-mnli	Anger	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Fear	50%
distilbert-base -uncased-mnli	Enjoyment	Sadness	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	41,7%
deberta-large -mnli-zero-cls	Fear	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	50%
Percentage correct (zsl)	0%	0%	0%	33,3%	0%	0%	33,3%	100%	33,3%	0%	0%	33,3%	
Zero-shot pipeline including 0.4 threshold													
bart-large-mnli	Anger	Enjoyment	Anger	Fear	Enjoyment	Enjoyment	Fear	Enjoyment	Fear	Sadness	Enjoyment	Enjoyment	25%
distilbert-base -uncased-mnli	Enjoyment	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	25%
deberta-large -mnli-zero-cls	Enjoyment	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Fear	Sadness	Enjoyment	Enjoyment	8,3%
Percentage correct (zsl + threshold 0.4)	0%	0%	0%	33,3%	33,3%	33,3%	66,7%	100%	66,7%	0%	0%	66,7%	
Zero-shot pipeline including 0.5 threshold													
bart-large-mnli	Anger	Enjoyment	Anger	Fear	Enjoyment	Enjoyment	Fear	Enjoyment	Fear	Anger	Enjoyment	Fear	33,3%
distilbert-base -uncased-mnli	Enjoyment	Sadness	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	41,7%
deberta-large -mnli-zero-cls	Enjoyment	Enjoyment	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Enjoyment	Sadness	Enjoyment	Enjoyment	25%
Percentage correct (zsl + threshold 0.5)	33,3%	0%	0%	0%	100%	100%	33,3%	100%	66,7%	33,3%	0%	100%	
Zero-shot pipeline including 0.6 threshold													
bart-large-mnli	Enjoyment	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Fear	Fear	Enjoyment	Fear	41,7%
distilbert-base -uncased-mnli	Enjoyment	Sadness	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	41,7%
deberta-large -mnli-zero-cls	Fear	Anger	Anger	Enjoyment	Sadness	Sadness	Fear	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	58,3%
Percentage correct (zsl + threshold 0.6)	33,3%	0%	0%	0%	100%	100%	0%	100%	100%	33,3%	0%	100%	

Table 6.5: Comparison of the results on excerpts of window size 7 with thresholds.

There are two models who obtain the best score, that is 58.3% compared to human annotation in a window size setting. One model is the GPT-3 model combined with a window of size 5, the other one is DistilBERT combined with window size 3. This can be seen in Figure 6.6 and Figure 6.7. it cannot be immediately inferred that one window size yields better results than the other, but the differences between models are clear. Where DistilBERT is the best performing model combined with window size 3, it yields the worse results in a window size 7 setting. This may be due

to the inability to process a vast amount of context words. In contrast, DeBERTa scores better the more context there is involved. The performance of all models combined with the different approaches can be observed in Figure 6.6, Figure 6.7, and Figure 6.8.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Fear	Fear	Fear	Fear	Fear	Sadness	Sadness	Enjoyment	Fear	Fear	Fear	Fear	
GPT-3	Sadness	Fear	Anger	Fear	Fear	Fear	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	58,3%
Cosine Similarity													
bart-large-mnli	Anger	Anger	Anger	Anger	Anger	Sadness	Sadness	Sadness	Fear	Sadness	Anger	Anger	25%
distilbert-base -uncased-mnli	Enjoyment	Fear	Fear	Fear	Fear	Enjoyment	Fear	Fear	Enjoyment	Sadness	Fear	Fear	50%
deberta-large -mnli-zero-cls	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Anger	Enjoyment	Sadness	Sadness	16,7%
Percentage correct (cos sim)	0%	33,3%	33,3%	33,3%	33,3%	66,7%	66,7%	0%	33,3%	0%	33,3%	33,3%	
Zero-shot pipeline													
bart-large-mnli	Anger	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Sadness	Sadness	Anger	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Sadness	Sadness	Enjoyment	25%
deberta-large -mnli-zero-cls	Sadness	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Ejoyment	Sadness	Fear	Fear	33,3%
Percentage correct (zsl)	0%	0%	0%	0%	0%	100%	33,3%	100%	0%	33,3%	66,7%	66,7%	
Percentage correct (gpt3 + cos sim + zsl)	0%	28,6%	14,3%	28,6%	28,6%	71,4%	42,9%	57,1%	14,3%	28,6%	57,1%	57,1%	

Table 6.6: Comparison of the results on excerpts of window size 3.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Sadness	Sadness	Fear	Fear	Sadness	Sadness	Sadness	Fear	Sadness	Fear	Fear	Fear	
GPT-3	Fear	Fear	Fear	Anger	Fear	Fear	Enjoyment	Fear	Enjoyment	Fear	Fear	Fear	41,7%
Cosine Similarity													
bart-large-mnli	Sadness	Anger	Anger	Anger	Anger	Anger	Enjoyment	Enjoyment	Enjoyment	Sadness	Anger	Anger	8,3%
distilbert-base -uncased-mnli	Fear	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Enjoyment	Enjoyment	Fear	Sadness	Enjoyment	Enjoyment	0%
deberta-large -mnli-zero-cls	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	41,7%
Percentage correct (cos sim)	66,7%	33,3%	0%	0%	33,3%	33,3%	33,3%	0%	0%	0%	0%	0%	
Zero-shot pipeline													
bart-large-mnli	Enjoyment	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Enjoyment	41,7%
deberta-large -mnli-zero-cls	Enjoyment	Anger	Anger	Fear	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Fear	Fear	41,7%
Percentage correct (zsl)	33,3%	33,3%	0%	33,3%	100%	100%	33,3%	0%	0%	33,3%	66,7%	66,7%	
Percentage correct (gpt3 + cos sim + zsl)	42,9%	28,6%	14,3%	14,3%	57,1%	57,1%	28,6%	14,3%	0%	28,6%	42,9%	42,9%	

Table 6.7: Comparison of the results on excerpts of window size 5.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Fear	Fear	Fear	Fear	Sadness	Sadness	Fear	Enjoyment	Enjoyment	Fear	Sadness	Fear	
GPT-3	Fear	Fear	Anger	Enjoyment	Enjoyment	Fear	Enjoyment	Enjoyment	Enjoyment	Anger	Enjoyment	Fear	41,7%
Cosine Similarity													
bart-large-mnli	Fear	Enjoyment	Sadness	Anger	Enjoyment	Sadness	Enjoyment	Anger	Anger	Fear	Sadness	Sadness	33,3%
distilbert-base -uncased-mnli	Fear	Fear	Sadness	Enjoyment	Fear	Sadness	Fear	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	58,3%
deberta-large -mnli-zero-cls	Sadness	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	33,3%
Percentage correct (cos sim)	66,7%	33,3%	0%	0%	33,3%	100%	33,3%	66,7%	33,3%	33,3%	66,7%	33,3%	
Zero-shot pipeline													
bart-large-mnli	Anger	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Fear	Enjoyment	Fear	50%
distilbert-base -uncased-mnli	Enjoyment	Sadness	Anger	Sadness	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	41,7%
deberta-large -mnli-zero-cls	Fear	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Enjoyment	Sadness	Enjoyment	Fear	50%
Percentage correct (zsl)	33,3%	0%	0%	0%	100%	100%	0%	100%	100%	33,3%	0%	100%	
Percentage correct (gpt3 + cos sim + zsl)	57,1%	28,6%	0%	0%	57,1%	85,7%	14,3%	85,7%	71,4%	28,6%	28,6%	71,4%	

Table 6.8: Comparison of the results on excerpts of window size 7.

In similar fashion to the single-sentence classification, there are excerpts who get an unanimous classification and excerpts that are a bit more divided when it comes to window size classification. The models and

the results of the user study are in agreement for the excerpt shown in Figure 6.1. It is unanimous considered to belong to the class *fear*.

Interviewer	**Child**
Hm-hm.	And I really needed to go to the toilet.
Okay, and then what happened next?	I went to the toilet by myself because mum couldn't leave my baby brother by himself.
Ah-ha.	And then I went to the toilet by myself.
Okay, and tell me everything that happened from when you went to the toilet.	I went into the toilet and I didn't lock the door because I didn't want to get stuck in.
Hm-hm.	And then the bad man came inside.

Figure 6.1: Excerpt from the user study with window size 5 where the models are in agreement with the human opinion.

There is not always a reason why one or both models are wrong; sometimes it is a matter of interpretation that differs between the models. In Figure 6.2, the models predict a different emotion based on based on the text given to them.

Interviewer	**Child**
Tell me more about the part where he put his hand in your undies.	He wrapped his hand around my doodle and then I said I didn't like it.
Hm-hm.	And then he laughed and took it out.
Okay, what happens next?	And then I went home.
Hm-hm, and was anyone else with you at uncle George?	No.
Hm-hm, and can you tell me more about the big boys game? Have you played it more than one time?	Yes.

Figure 6.2: Excerpt from the user study with window size 5 for which both GPT-3 and the human raters agreed that this should be classified as *fear*, while the BART model classified its as *anger*.

However, there are also instances where the models make different classifications than human participants. This happens with the conversation shown in Figure 6.3. The models only see the child's responses, and thus, classify this excerpt as *enjoyment*. The participants, on the other hand, see the whole conversation and classified it as *fear* instead. Due to these circumstances, the decision was made to experiment with providing the model with both the questions from the interviewer and the child's response. We will talk about these result in the next subsection.

Interviewer	Child
Okay, so no one else was home. Where was everyone else?	Work.
At work, okay, and why was Mark there?	To play.
To play, okay. Okay, so you've told me quite a bit. Let me just doublecheck that I've got all of that right. So Mark was with you and you were playing Nintendo and you stood up and broke the controller. Then he got angry at you for that so you went to the garden and got a stick and then he hit you with it lots of times and then after that, he just left and he told you not to tell anyone.	Yes.

Figure 6.3: Excerpt from the user study with window size 3 for which models were not in agreement with the human raters.

6.2.3 Window size excerpts with questions from the interviewer

As a result of the disappointing results from the threshold experiments, we decided not to test the threshold in combination with the questions from the interviewers.

Due to the example in Figure 6.3, it was worth researching whether better results would be obtained if the interviewers' questions were also included in the classification. The classification was made on the entire sequence of question and answer as a whole, and not as two separate classification concatenated.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Fear	Fear	Fear	Fear	Fear	Sadness	Sadness	Enjoyment	Fear	Fear	Fear	Fear	
GPT-3	Sadness	Fear	Anger	Fear	Fear	Anger	Sadness	Fear	Anger	Fear	Fear	Fear	58,3%
Cosine Similarity													
bart-large-mnli	Anger	Anger	Anger	Anger	Anger	Anger	Anger	Anger	Anger	Anger	Anger	Anger	0%
distilbert-base -uncased-mnli	Enjoyment	Fear	Anger	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	16,7%
deberta-large -mnli-zero-cls	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Enjoyment	Sadness	Sadness	16,7%
Percentage correct (cos sim)	0%	33,3%	0%	0%	0%	33,3%	33,3%	33,3%	0%	0%	0%	0%	
Zero-shot pipeline													
bart-large-mnli	Sadness	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Anger	Enjoyment	Fear	Anger	25%
distilbert-base -uncased-mnli	Sadness	Sadness	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Anger	Sadness	Fear	Anger	25%
deberta-large -mnli-zero-cls	Fear	Anger	Anger	Enjoyment	Sadness	Sadness	Enjoyment	Enjoyment	Anger	Sadness	Fear	Fear	41,7%
Percentage correct (zsl)	33,3%	0%	0%	0%	0%	100%	0%	100%	0%	0%	100%	33,3%	
Percentage correct (gpt3 + cos sim + zsl)	14,3%	28,6%	0%	14,3%	14,3%	57,1%	28,6%	57,1%	0%	14,3%	57,1%	28,6%	

Table 6.9: Comparison of the results on excerpts of window size 3 including the interviewers questions.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Sadness	Sadness	Fear	Fear	Sadness	Sadness	Sadness	Fear	Sadness	Fear	Fear	Fear	
GPT-3	Fear	Fear	Fear	Anger	Fear	Sadness	Anger	Sadness	Sadness	Sadness	Fear	Fear	41,7%
Cosine Similarity													
bart-large-mnli	Anger	Anger	Anger	Anger	Anger	Enjoyment	Anger	Anger	Anger	Anger	Anger	Enjoyment	0%
distilbert-base -uncased-mnli	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	0%
deberta-large -mnli-zero-cls	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	41,7%
Percentage correct (cos sim)	33,3%	0%	0%	0%	33,3%	33,3%	33,3%	0%	33,3%	0%	0%	0%	
Zero-shot pipeline													
bart-large-mnli	Anger	Anger	Enjoyment	Fear	Sadness	Sadness	Anger	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	41,7%
distilbert-base -uncased-mnli	Enjoyment	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Sadness	Fear	Enjoyment	Sadness	Enjoyment	33,3%
deberta-large -mnli-zero-cls	Sadness	Enjoyment	Fear	Fear	Sadness	Sadness	Anger	Enjoyment	Enjoyment	Enjoyment	Fear	Fear	58,3%
Percentage correct (zsl)	33,3%	33,3%	33,3%	66,7%	100%	100%	33,3%	0%	0%	0%	66,7%	66,7%	
Percentage correct (gpt3 + cos sim + zsl)	28,6%	14,3%	28,6%	28,6%	57,1%	71,4%	28,6%	0%	28,6%	0%	42,9%	42,9%	

Table 6.10: Comparison of the results on excerpts of window size 5 including the interviewers questions.

	Story_1	Story_2	Story_3	Story_4	Story_5	Story_6	Story_7	Story_8	Story_9	Story_10	Story_11	Story_12	Percentage correct
User study	Fear	Fear	Fear	Fear	Sadness	Sadness	Fear	Enjoyment	Enjoyment	Fear	Sadness	Fear	
GPT-3	Fear	Anger	Anger	Fear	Sadness	Fear	Anger	Enjoyment	Enjoyment	Fear	Sadness	Fear	66,7%
Cosine Similarity													
bart-large-mnli	Anger	Anger	Anger	Anger	Anger	Enjoyment	Sadness	Anger	Anger	Anger	Anger	Anger	0%
distilbert-base -uncased-mnli	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	Enjoyment	16,7%
deberta-large -mnli-zero-cls	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	Sadness	25%
Percentage correct (cos sim)	0%	0%	0%	0%	33,3%	33,3%	0%	33,3%	33,3%	0%	33,3%	0%	
Zero-shot pipeline													
bart-large-mnli	Enjoyment	Enjoyment	Anger	Anger	Enjoyment	Sadness	Fear	Enjoyment	Enjoyment	Fear	Sadness	Fear	58,3%
distilbert-base -uncased-mnli	Sadness	Anger	Anger	Sadness	Sadness	Sadness	Enjoyment	Sadness	Sadness	Sadness	Sadness	Fear	33,3%
deberta-large -mnli-zero-cls	Fear	Anger	Anger	Enjoyment	Enjoyment	Sadness	Fear	Enjoyment	Enjoyment	Sadness	Sadness	Fear	58,3%
Percentage correct (zsl)	33,3%	0%	0%	0%	33,3%	100%	66,7%	66,7%	66,7%	33,3%	100%	100%	
Percentage correct (gpt3 + cos sim + zsl)	28,6%	0%	0%	14,3%	42,9%	57,1%	28,6%	57,1%	57,1%	28,6%	71,4%	57,1%	

Table 6.11: Comparison of the results on excerpts of window size 7 including the interviewers questions.

Whenever we look at the classifications where the interviewers' question was included, we can see that the cosine similarity model does not perform well. When we take a closer look, we can see that the cosine similarities are all very close together, meaning that the models have a problem to classify with confidence. This is probably the case because there is too much context information for these models to actually classify it as one of the four emotions portrayed. Even though the hypothesis was that the results would improve, they did not.

The complete overview of classifications per model and approach can be found on the GitHub repository belonging to this thesis.

Chapter 7

Discussion

There were also promising initial results concerning the emotional component. However, it is essential that more annotated data gets analysed and compared to the output of the existing models. When this is done, and the best model has been chosen, we can incorporate the suitable model into our emotion pipeline. Subsequently, we can start implementing the pipeline in both the auditory and the visual components. Users will evaluate this implementation to determine what works and what does not.

The main limitations of the current pilot study are the small size of the sample. Although pilot studies usually have small samples, results from current research may not be generalised to a broader population.

The threshold did not seem to work as well as we expected due to many miss-classifications for the *enjoyment* emotion. It is worth experimenting to see whether it would improve if we were to ignore a threshold for *enjoyment* but keep it for the other emotions. This may yield better results since the prevailing emotion in child maltreatment is generally more sad than happy. The window size approach seems like a good balance between the single-sentence and the whole-story-so-far classification, where there is too little and too much context available respectively. The optimal window size depends on the model and the approach.

Chapter 8

Conclusion and future work

8.1 Summary

This thesis exposes the current performance of language models and chatbot frameworks associated, on the one hand, with interview environments that mimic abused, neglected, and traumatised children and, on the other hand, with sentiment analysis.

We show that a non-fine-tuned approach is not satisfactory when it comes to the pre-trained GPT-2 and DialoGPT models. Therefore, we developed an initial chatbot using the RASA framework. Not only is this a working example of Talking Child-Avatar, but we also created and implemented a sentiment analysis pipeline. The pipeline is a tool to extract the emotion of both the user input and the bot's response. Other system components will use the emotion classification output as their input. The audio and visual parts depend on the emotions to alter their output to portray the correct characteristics corresponding to the specified emotion.

We conducted multiple user studies to see how well humans can extract emotions from mock interviews. These user studies show excerpts of the mock-interview transcripts. Every participant has to classify an excerpt as one of the predefined emotions. It showed that humans had difficulties to reach a consensus when they had to choose one out of seven possible emotions. The unity of the answers improved when the number of possible emotions was reduced from seven to four. The extraction gets compared between single-sentence excerpts and story excerpts with different window sizes. The windows are either of size 3, 5, or 7.

We studied multiple ready-available language models to use for sentiment classification in a pure textual setting without the help of video material to substantiate the decisions. The models are implemented using two different settings, a cosine similarity approach and within the Huggingface zero-shot pipeline. GPT-3 was evaluated through their API. The zero-shot pipeline outperformed the cosine similarity approach in almost all situations. It also performed best on single-sentence excerpts. However, GPT-3 was the best model when context comes into play, which is of great importance for this research project. We can conclude that emotion extraction based on the whole story does not work as there is

too much context present. Single-sentence classification is not suitable for our project as context is essential when it comes to sentiment analysis. The different window sizes achieve different results based on the different models as some models require more context than other to make a correct sentiment classification. The inclusion of the interviewers' questions and the threshold approach often worsened the results in contrast to our initial beliefs.

8.2 Revisiting the research questions

8.2.1 Chatbot framework

The first research question we want to answer is: **What chatbot frameworks and language models are currently available? What are their pros and cons when it comes to this research project?**

As of today, there are many language models and chatbot frameworks available, all with their own pros and cons. The most promising language models for our project were GPT-2 and GPT-3. Where GPT-2 was not able to live up to the expectations, the possibility of future success for GPT-3 is great. However, we seem to miss more annotated data to either fine-tune these models or train our own model from scratch. Training our own model would be a worthy approach as this use case is quite specific, and the available models out there are trained on adult language instead of children's language. Furthermore, they seem to miss the nuances of the setting of child abuse and maltreatment. We branched out to the chatbot framework RASA as this is not yet available. RASA is the perfect intermediate solution, but there are definitely flaws. Firstly, it is a lot of manual work to convert transcripts into RASA accepted formats. Secondly, RASA has a problem when it comes to holding a conversation based on small talk. The modelling of small-talk and other questions is possible; however, this leads to too many intent definitions, which negatively impacts the performance of the intent classification model, resulting in poorer chatbots. Last but not least, it is also not able to generate its own storyline based on the available transcripts. Therefore, it is not the desired solution since one of the main objectives of the project is to generate new storylines every time so that trainees are subject to new material every time.

8.2.2 Sentiment classification

The next two research questions that we want to answer are: How well can state-of-the-art models extract emotions from available transcripts compared to human annotation? How can we improve this emotion classification?

Before we can answer this question, it is important to take another look at the human annotations from the user studies. Evidently, it is challenging for humans to reach a consensus when classifying data excerpts with

seven emotions as the options. There was unanimous agreement about only one of the twenty-one questions, that one being a single sentence excerpt. Although there was a clear winner for eleven questions, the ICC based on average measures was low in 7 dimensions, scoring only 0.537 on average. When we reduced the number of emotions to four, the participants were better at agreeing on the best-fitting class. Reducing the number of categories, resulting in more clear winners than before. The ICC also increased from 0.537 to 0.788: significantly higher with four dimensions than with seven. Therefore, we chose to perform the final experiments with only the subset of four emotions.

Furthermore, there are different strategies that have been used to experiment with. The first one is the single-sentence prediction. This one is accurate when compared to human annotation but not precise enough since it misses important context. In contrast, the whole story has too much context, resulting in results that are also not accurate. The sliding window appeared to offer the perfect solution to this problem. Using a window consisting of between 3 and 7 sentences was expected to be ideal.

In theory, both the whole-story prediction and the sliding window can be used with the strategy to implement a threshold. However, the threshold approach does not not yield the desired results. The hypothesis was that it would help spot sudden substantial emotional changes in the story, for example, if a child started crying. However, a restart of the window often occurs with sentences classified as *enjoyment*. Consequently, the model predictions got worse due to the implementation of the threshold. The same was the case when we included the interviewers' questions into the sequence to be classified. It seemed that there was too much redundant context that often worsened the results. So to conclude, the three different strategies were single-sentence prediction, whole-story prediction, and a sliding window.

Not only did we experiment with the number of emotions and different strategies, we also used different models in combination with all the different strategies. The different models are the GPT-3 from OpenAI, a DeBERTa model from Narsil, a DistilBERT model from Typeform, and a BART model from Facebook. All these models yielded various results combined with either the cosine similarity or the zero-shot attempt, and also in combination with the different strategies. Generally, we can say that GPT-3 performed best out of the models we tests, especially in combination with either window size 3 excluding the interviewer's questions or with window size 7 including the interviewer's questions. It did not perform best for single-sentence classification, that was BART, but we require the influence of context for the sentiment classification of the Talking Child-Avatar.

The aim of the Talking Child-Avatar is to extract emotions from child-police mock-interview transcripts to be used in different components of the Talking Child-Avatar. This research paper shows that there are multiple approaches to extracting these emotions based on textual data, each with its own advantages and disadvantages. Furthermore, the conducted user studies give us valuable insights into the way that humans classify

emotions in both a seven- and four-dimension setting. In addition, the creation of the sentiment pipeline enables the extraction of these emotions and simultaneously supplies these as input for the visual and audio parts of the Talking Child-Avatar.

All code and results can be found at https://github.com/MyrtheLammerse/thesis.

8.3 Future work

Future studies with larger sample sizes may provide important information about variables such as attrition, participation, and effectiveness. This allows for a better understanding of effectiveness throughout various populations, such as older or younger participants. It is important to have more annotated data available in order to assess the performance of the different models and approaches. The extra data can also be used to fine-tune the available models and go from a zero-shot to a few-shot setting. It is also worth researching whether the classification will improve if the question and answer are classified separately and then concatenated instead of classifying it as a whole as has been done in this research. Furthermore, fine-tuning of the data by, for example, stop-word removal may be of a positive influence. Furthermore, an improvement would be to allow multi-class classification where not just one emotion is considered as the truth, but that it can be a mix of many. It is necessary to get more data with more different emotions annotated to make this work.

However, now, the most significant improvements would be made if there were annotated text data available in combination with the corresponding video input. Not only can we use this to predict, but it is crucial in cases where numbing plays a role.

So far, we have not paid much attention to the creation of personas. When the right model has been chosen, and the emotion extraction is nearly perfect, this needs to be a point of attention. The personas need to be implemented to make it as real as possible as not every child expresses emotions to the same degree; this is also a way to incorporate numbing into the system.

The GPT3 platform also has a translation option for future improvements. Therefore, it will be possible to use this system in multiple languages.

When all this is complete, the complete chatbot, with the integrated emotion component, should be evaluated by experts in the field of investigative interviewing. Experts can be, for example, police officers, employees from child protective services, Ph.D. students, and professors from the social faculty with experience in working with children.

In conclusion, here is still a lot of work to be done in order to create the Talking Child-Avatar. RASA is a good solution for now, but we need more annotated data to fine-tune the GPT-3 model. Using GPT-3 will allow us to have conversations in a way that is not possible with RASA. Small talk and

out of scope questions are hard to process within the RASA environment. Therefore, we need a language model that is able to generate language and is not dependent on predefined storylines. The other beneficial aspect of generative aspect is that the user will be subjected to a new story every time they use the Talking Child-Avatar. These changes should be investigated as a promising way to improve the work done in light of the chatbot framework research question.

Furthermore, more annotated needs to be obtained, preferably in combination with the video, in order to improve and assess the sentiment classification results of the existing models. These alterations should be explored for possible improvements in light of the research question about sentiment classification.

Appendix A

Appendix A - user studies

A.1 User study 1

The form should be anonymous. **Show more**

User study - sentiment chatbot

Mandatory fields are marked with a star *

Welcome and thank you for participating in this user study.

The aim of this research is to develop a new digital interview-training program drawing on expertise in developmental psychology and artificial intelligence. We need your help to predict and model emotions as true as possible.

You will first get asked a couple of personal questions and then you will get to see a few excerpts from police interviews that you will have to classify as a certain emotion, more information about that later.

Thank you for taking your time in assisting me with this research. Under no circumstances are you obliged to answer any of the questions. However, doing so will greatly assist me in completing my research and enhancing the understanding of emotions. The data collected will remain confidential and used solely for academic purposes.

- Myrthe Lammerse, Language Technology student at the University of Oslo.

Trigger warning:

This survey includes readings around topics such as sexual assault, domestic violence, physical violence, child abuse, and child harassment. I acknowledge that this content may be difficult. I also encourage you to care for your safety and well-being while filling in the survey.

How old are you? *

- ◯ < 18
- ◯ 18 - 25
- ◯ 26 - 35
- ◯ 36 - 45
- ◯ 46 - 55
- ◯ 56 - 65
- ◯ > 65

With which gender do you identify?

- ◯ Female
- ◯ Male
- ◯ Non-binary
- ◯ Other
- ◯ Prefer not to say

What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

- ◯ No schooling completed
- ◯ High school
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ Doctoral degree
- ◯ Other

We will present you with a couple of text excerpts that you will have to classify with an emotion. The excerpts are from interviews between police officers and children. You will have to classify the example as one of the possible emotions.

The possible emotions are enjoyment, surprise, fear, sadness, anger, disgust, and contempt.

The emotions allude to the following:

Enjoyment is, for many, the most desirable of the seven universal emotions, typically arising from connection or sensory pleasure. It describes a family of pleasurable states, everything from peace to ecstasy.

Surprise arises when we encounter sudden and unexpected sounds or movements. Its function is to focus our attention on determining what is happening and whether or not it is dangerous.

Fear arises with the threat of harm, either physical, emotional, or psychological, real or imagined. While traditionally considered a "negative" emotion, fear actually serves an important role in keeping us safe as it mobilizes us to cope with potential danger.

Sadness results from the loss of someone or something important. What causes us sadness varies greatly based on personal and cultural notions of loss. Sadness describes the range of emotional states we can experience containing everything from mild disappointment to extreme despair and anguish.

Anger arises when we are blocked from pursuing a goal and/or treated unfairly. communicates anything from mere dissatisfaction to threats.

Disgust arises as a feeling of aversion towards something offensive. It contains a range of states with varying intensities from mild dislike to intense loathing. All states of disgust are triggered by the feeling that something is averse, repulsive and/or toxic.

Contempt is the feeling of dislike for and superiority (usually morally) over another person, group of people, and/or their actions.

It's important to note that *contempt* is related to but different from *disgust*. While both contempt and disgust can be directed toward people, and their actions, disgust can also be aroused by objects that are aversive to the senses (taste, smell, sight, sound, touch). Additionally, contempt includes the feeling of superiority over the target of contempt, whereas one doesn't necessarily feel superior to the person/thing that disgusts them.

If you are not 100% sure, choose the option that is most applicable. There is no right or wrong answer.

We will start with single sentence examples.

What emotion best fits the following sentence?

"Just rude."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"Because I was scared I would get stuck inside."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"
He said that he had a fun game to play and only the big boys could play it."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"No."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"Sometimes it's funny."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"It was big and yuck."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"And then he ran away and then an ambulance came."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"Just in the art room."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"And we always have a big person to look after us."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"So I touched it."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"Yeah."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"He said that he wished he didn't live with us anymore."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

What emotion best fits the following sentence?

"I was on a chair."

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

Now you will get examples that span multiple sentences. Base your answer on all the answers the child has given.

Interviewer	Child
Do you know if the people on the computer where adults or children?	I think they were children.
And what makes you think they were children?	They were little.
And how many people did you see on the computer?	Two.
And what happened while you were looking at the people on the computer?	He said that he had to play a game with me.
Hm-hm.	And he put his hand in my undies.

What emotion best fits the child's answers?

- Enjoyment
- Surprise
- Fear
- Sadness
- Anger
- Disgust
- Contempt

Interviewer	Child
Okay, and what happened next?	He said that he has to help.
Yeah.	And he put his hand on my privates.
And then what happened?	Then I just started crying.
Hm-hm.	And he told me to be quiet.
Yeah.	And then we heard someone come into the toilet

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

Interviewer	Child
What do you call Janet's Grandfather?	I call him Brian
Brian?	Yeah
Okay, so tell me all about Brian	He's really tall and he's a bit fat and sometimes he looks after us
Okay and I've heard that you don't like going to Janet's house. Tell me why you don't like that.	Well, when we're there, he touches us on our girlie bits.
He touches you on your girlie bits? Can you tell me what you mean by your girlie bits?	Yes.

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

Interviewer	Child
Hi Sophie. Tell me everything you've come here to talk to me about today.	About my dad.
Okay, and what about your dad?	His stick hurt me.
Okay, and what do you mean when you say stick?	His private.
And when you say private, what does that mean?	It's what he uses to do a wee.
Okay, all right, and where did this happen, Sophie?	At his house.

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

Interviewer	Child
Hm-hm	And then Janet told him off because she saw what he did.
Hm-hm. What happened next?	Janet's pop just got out of the spa.
Hm-hm.	And then Janet's mum came home.
Hm-hm, and then what happened?	She was mad at us because she said that we weren't allowed to go in the spa without her but we just said sorry.
Hm-hm.	And she gave us some party pies.

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

Interviewer	Child
Did you ride your bike there?	Yeah.
Ah, and what's at the park?	There's a slide and some swings and a tunnel.
Ah, and Isabelle, do you go there very often?	Yeah.
How many times have you been there?	Lots of times.
Ah, and what was this day that you went there with your brother? Sorry, I'll say that again. What day was it that you went with your brother there, Isabelle?	On the weekend.

Interviewer	Child
Okay, and where were they?	They were in bed.
Oh okay, and you said that your cousin didn't say anything to you?	Yeah.
And you didn't say anything to him?	Yeah.
But you did tell your mum?	Yeah.
So Sarah, can you tell me what you've come here to talk to me about today?	Yeah.

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

Interviewer	Child
About the bad man. So tell me everything about the bad man, start from the beginning.	He was smelly and he was quite tall, and he had long brown hair.
Okay, so he was smelly, he was quite tall, and he had long brown hair.	Yeah.
Okay, so what happened when you saw the bad man?	I was in the toilet at Kmart and he came in.
In the toilet at Kmart, and then he came in. And then what happened?	And then he opened the door and he told me that he had to help me wipe.
He opened the door and told you that he had to help you wipe?	Yeah, and I told him I didn't need any help.

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

Interviewer	Child
Hm-hm.	But I didn't lock the door because I didn't want to get stuck inside.
Didn't want to get stuck inside, yeah, what happened then?	Then the bad man came in.
What happened then?	He was wearing a big green coat and he was really tall.
Hm-hm. What happened then?	And then he put his hand on my mini.
Tell me what you use your mini for.	To go wees.

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Surprise
- ◯ Fear
- ◯ Sadness
- ◯ Anger
- ◯ Disgust
- ◯ Contempt

Send

Responsible for the form: **myrthel@uio.no**.

A.2 User study 2

The form should be anonymous. **Show more**

User study - basic emotions sentiment chatbot

Mandatory fields are marked with a star *

Welcome and thank you for participating in this user study.

The aim of this research is to develop a new digital interview-training program drawing on expertise in developmental psychology and artificial intelligence. We need your help to predict and model emotions as true as possible.

You will first get asked a couple of personal questions and then you will get to see a few excerpts from police interviews that you will have to classify as a certain emotion, more information about that later.

Thank you for taking your time in assisting me with this research. Under no circumstances are you obliged to answer any of the questions. However, doing so will greatly assist me in completing my research and enhancing the understanding of emotions. The data collected will remain confidential and used solely for academic purposes.

- Myrthe Lammerse, Language Technology student at the University of Oslo.

Trigger warning:

This survey includes readings around topics such as sexual assault, domestic violence, physical violence, child abuse, and child harassment. I acknowledge that this content may be difficult. I also encourage you to care for your safety and well-being while filling in the survey.

How old are you? *

- ◯ < 18
- ◯ 18 - 25
- ◯ 26 - 35
- ◯ 36 - 45
- ◯ 46 - 55
- ◯ 56 - 65
- ◯ > 65

With which gender do you identify?

- ◯ Female
- ◯ Male
- ◯ Non-binary
- ◯ Other
- ◯ Prefer not to say

What is the highest degree or level of school you have completed? If currently enrolled, highest degree received.

- ◯ No schooling completed
- ◯ High school
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ Doctoral degree
- ◯ Other

We will present you with a couple of text excerpts that you will have to classify with an emotion. The excerpts are from interviews between police officers and children. You will have to classify the example as one of the possible emotions.

The possible emotions are enjoyment, fear, sadness, and anger.

The emotions allude to the following:

Enjoyment is, for many, the most desirable of the seven universal emotions, typically arising from connection or sensory pleasure. It describes a family of pleasurable states, everything from peace to ecstasy.

Fear arises with the threat of harm, either physical, emotional, or psychological, real or imagined. While traditionally considered a "negative" emotion, fear actually serves an important role in keeping us safe as it mobilizes us to cope with potential danger.

Sadness results from the loss of someone or something important. What causes us sadness varies greatly based on personal and cultural notions of loss. Sadness describes the range of emotional states we can experience containing everything from mild disappointment to extreme despair and anguish.

Anger arises when we are blocked from pursuing a goal and/or treated unfairly. communicates anything from mere dissatisfaction to threats.

If you are not 100% sure, choose the option that is most applicable. There is no right or wrong answer.

We will start with single sentence examples.

What emotion best fits the following sentence?

"Just rude."

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

What emotion best fits the following sentence?

"

He said that he had a fun game to play and only the big boys could play it."

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

What emotion best fits the following sentence?

"Because I was scared I would get stuck inside."

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

What emotion best fits the following sentence?

"No."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"It was big and yuck."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"Sometimes it's funny."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"Just in the art room."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"And then he ran away and then an ambulance came."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"And we always have a big person to look after us."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"Yeah."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"I was on a chair."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence?

"So I touched it."

- Enjoyment
- Fear
- Sadness
- Anger

Now you will get examples that span multiple sentences. Base your answer on all the answers the child has given.

Interviewer	Child
Do you know if the people on the computer where adults or children?	I think they were children.
And what makes you think they were children?	They were little.
And how many people did you see on the computer?	Two.
And what happened while you were looking at the people on the computer?	He said that he had to play a game with me.
Hm-hm.	And he put his hand in my undies.

What emotion best fits the following sentence?

"He said that he wished he didn't live with us anymore."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay, and what happened next?	He said that he has to help.
Yeah.	And he put his hand on my privates.
And then what happened?	Then I just started crying.
Hm-hm.	And he told me to be quiet.
Yeah.	And then we heard someone come into the toilet

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
What do you call Janet's Grandfather?	I call him Brian
Brian?	Yeah
Okay, so tell me all about Brian	He's really tall and he's a bit fat and sometimes he looks after us
Okay and I've heard that you don't like going to Janet's house. Tell me why you don't like that.	Well, when we're there, he touches us on our girlie bits.
He touches you on your girlie bits? Can you tell me what you mean by your girlie bits?	Yes.

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay, and where were they?	They were in bed.
Oh okay, and you said that your cousin didn't say anything to you?	Yeah.
And you didn't say anything to him?	Yeah.
But you did tell your mum?	Yeah.
So Sarah, can you tell me what you've come here to talk to me about today?	Yeah.

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Did you ride your bike there?	Yeah.
Ah, and what's at the park?	There's a slide and some swings and a tunnel.
Ah, and Isabelle, do you go there very often?	Yeah.
How many times have you been there?	Lots of times.
Ah, and what was this day that you went there with your brother? Sorry, I'll say that again. What day was it that you went with your brother there, Isabelle?	On the weekend.

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hi Sophie. Tell me everything you've come here to talk to me about today.	About my dad.
Okay, and what about your dad?	His stick hurt me.
Okay, and what do you mean when you say stick?	His private.
And when you say private, what does that mean?	It's what he uses to do a wee.
Okay, all right, and where did this happen, Sophie?	At his house.

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm	And then Janet told him off because she saw what he did.
Hm-hm. What happened next?	Janet's pop just got out of the spa.
Hm-hm.	And then Janet's mum came home.
Hm-hm, and then what happened?	She was mad at us because she said that we weren't allowed to go in the spa without her but we just said sorry.
Hm-hm.	And she gave us some party pies.

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
About the bad man. So tell me everything about the bad man, start from the beginning.	He was smelly and he was quite tall, and he had long brown hair.
Okay, so he was smelly, he was quite tall, and he had long brown hair.	Yeah.
Okay, so what happened when you saw the bad man?	I was in the toilet at Kmart and he came in.
In the toilet at Kmart, and then he came in. And then what happened?	And then he opened the door and he told me that he had to help me wipe.
He opened the door and told you that he had to help you wipe?	Yeah, and I told him I didn't need any help.

What emotion best fits the child's answers?

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm.	But I didn't lock the door because I didn't want to get stuck inside.
Didn't want to get stuck inside, yeah, what happened then?	Then the bad man came in.
What happened then?	He was wearing a big green coat and he was really tall.
Hm-hm. What happened then?	And then he put his hand on my mini.
Tell me what you use your mini for.	To go wees.

What emotion best fits the child's answers?

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Send

Responsible for the form: **myrthel@uio.no**.

Terms and conditions

A.3 User study 3

The form should be anonymous. **Show more**

User study - basic emotions sentiment chatbot - extended

Mandatory fields are marked with a star *

Welcome and thank you for participating in this user study.

The study will take about 15-25 minutes.

The aim of this research is to develop a new digital interview-training program drawing on expertise in developmental psychology and artificial intelligence. We need your help to predict and model emotions as true as possible.

You will first get asked a couple of personal questions and then you will get to see a few excerpts from police interviews that you will have to classify as a certain emotion, more information about that later.

Under no circumstances are you obliged to answer any of the questions. However, doing so will greatly assist me in completing my research and enhancing the understanding of emotions. The data collected will remain confidential and used solely for academic purposes.

- Myrthe Lammerse, Language Technology student at the University of Oslo.

Trigger warning:

This survey includes readings around topics such as sexual assault, domestic violence, physical violence, child abuse, and child harassment. I acknowledge that this content may be difficult. I also encourage you to care for your safety and well-being while filling in the survey.

How old are you? *

- ◯ < 18
- ◯ 18 - 25
- ◯ 26 - 35
- ◯ 36 - 45
- ◯ 46 - 55
- ◯ 56 - 65
- ◯ > 65

With which gender do you identify? *

- ◯ Female
- ◯ Male
- ◯ Non-binary
- ◯ Other
- ◯ Prefer not to say

What is the highest degree or level of school you have completed? If currently enrolled, highest degree received. *

- ◯ No schooling completed
- ◯ High school
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ Doctoral degree
- ◯ Other

We will present you with a couple of text excerpts that you will have to classify with an emotion. The excerpts are from interviews between police officers and children. You will have to classify the example as one of the possible emotions.

The possible emotions are enjoyment, fear, sadness, and anger.

The emotions allude to the following:

Enjoyment is, for many, the most desirable of the seven universal emotions, typically arising from connection or sensory pleasure. It describes a family of pleasurable states, everything from peace to ecstasy.

Fear arises with the threat of harm, either physical, emotional, or psychological, real or imagined. While traditionally considered a "negative" emotion, fear actually serves an important role in keeping us safe as it mobilizes us to cope with potential danger.

Sadness results from the loss of someone or something important. What causes us sadness varies greatly based on personal and cultural notions of loss. Sadness describes the range of emotional states we can experience containing everything from mild disappointment to extreme despair and anguish.

Anger arises when we are blocked from pursuing a goal and/or treated unfairly. It communicates anything from mere dissatisfaction to threats.

If you are not 100% sure, choose the option that is most applicable. There is no right or wrong answer.

We will start with single sentence examples.

What emotion best fits the following sentence? *

"Then he tried to push it in and it really hurt."

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

What emotion best fits the following sentence? *

"And then I got taken to the principal."

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

What emotion best fits the following sentence? *

"He starts shouting and he was swearing and then he hit Mum."

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

What emotion best fits the following sentence? *

"He hurt mum and me."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence? *

"Because I was hiding there."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence? *

"
We watched a movie and then we got some ice cream and then we went to bed."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence? *

"It really hurt."

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the following sentence? *

"He told us that we had to go into the cubicle one at a time."

- Enjoyment
- Fear
- Sadness
- Anger

Now you will get examples that span multiple sentences. Base your answer on all the answers the child has given.

Interviewer	Child
Okay, so when you say he wiped. Can you describe to me how he did that?	He just wiped the wee.
Hm-hm. Did he use anything?	Yeah.
What was that?	Just his fingers.
Okay, and then what happened?	And then I started crying.
Hm-hm.	And then he ran away.
He ran away?	Yeah.
Okay, so you mentioned before that he said he had to help you. Can you tell me more about that/	Yeah.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Did he say anything else to you?	Yeah.
Tell me what he said.	He just said that his job was to help all the little girls in the toilet.
Yeah.	And he said that he has to help me.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay, so no one else was home. Where was everyone else?	Work.
At work, okay, and why was Mark there?	To play.
To play, okay. Okay, so you've told me quite a bit. Let me just doublecheck that I've got all of that right. So Mark was with you and you were playing Nintendo and you stood up and broke the controller. Then he got angry at you for that so you went to the garden and got a stick and then he hit you with it lots of times and then after that, he just left and he told you not to tell anyone.	Yes.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay, and what happened next?	And then he showed me some rude people on the computer.
On the computer?	Yeah.
Okay, and this is in the lounge room of his house?	Yeah.
Okay, when you say some rude people. What were the pictures?	Naked people.
Naked, and what were they doing?	Touching each other's privates.
Okay, so let me just go over what you told me. You said that you were in his lounge room and he was on his computer. And where were you sitting?	I wasn't sitting.
Okay, so where were you when he was on his computer? Where were you?	Standing next to him.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm. What happened after that?	And then we went back to play on the playground.
Hm. Did anything else happen to you while you were in the toilet with Sam?	No.
You said there were other boys there?	Yes.
What happened with the other boys?	They were just watching.
Hm-hm. What were they doing when they were watching?	They were just laughing.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
So tell me how old Mark is.	He's 17.
Sarah, tell me the last time you saw Mark.	I think it was last week.
Tell me, Sarah. Tell me about when you last saw Mark. Where you were and things like that.	I came home from school and he was looking after me and he started saying nasty things to me.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm.	And when I got there, we were playing under the sprinkler.
Okay. So you were playing under the sprinkler? And was that with Janet?	Yeah.
And what happened next?	Then her grandfather came out and started spraying us with the hose.
Okay, and what happened after that?	And then he asked us if we wanted to go in the spa.
Okay, so he asked you if you wanted to go into the spa. And then what happened after that?	We're only allowed to go in the spa with an adult.
Hm-hm.	So the three of us went in together.
Okay, so what happened after you went into the spa with Janet and her grandfather?	Janet and I were splashing each other.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm.	And I really needed to go to the toilet.
Okay, and then what happened next?	I went to the toilet by myself because mum couldn't leave my baby brother by himself.
Ah-ha.	And then I went to the toilet by myself.
Okay, and tell me everything that happened from when you went to the toilet.	I went into the toilet and I didn't lock the door because I didn't want to get stuck in.
Hm-hm.	And then the bad man came inside.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
You don't live with him. So did something happen to you on the weekend?	Yes.
Can you tell me more about that?	Yes.
So what sort of thing happened?	He hurt me.
He hurt you. And can you tell me how he hurt you?	Yes.
And can you explain?	Yes.
Hm.	He just hurt me.
Pardon?	He just hurt me.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay, that's fine, thank you. And whereabouts where you in the house when daddy's stick hurt you?	In my bed.
In your bed, okay. And was anybody else there?	No.
No, okay, I see. And what were you doing on your bed?	I was going to sleep.
Okay, what happened then when you were going to sleep?	Then he came in and he said he had a present for me.
Oh, okay. He said he has a present, okay. And what was dad's present?	I wasn't really a present.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay, so when pop grabbed your girlie bits, was it for a short time or a long time?	For a short time.
For a short time. Okay. And did pop say anything?	No.
Okay, so tell me more about when pop grabbed you on the girlie bits.	He was just sitting next to me and then he put his hand down and I could feel it there.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
And what happens after that?	And then I played outside.
You play outside. Okay, you've done a great job in talking to me today. You're doing really well. Tell me just form the beginning when he follows you into the cupboard. Just tell me that bit when he follows you into the cupboard. So the bit when he goes first.	He pulls his pants down.
Yeah, what then?	He showed me his willy.
So he follows you into the cupboard and then he pulls his pants down and shows you his willy? Have I got that right?	Hm-hm.
Okay, and what then?	Then I have to pull my dress up.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay, you said that that happened more than one time. Did it happen more than one time on different days or more than one time on the same day?	On different days.
Okay, tell me about another time that you can remember.	I woke up and his dirty spot was in his hand and then we heard a noise.
Okay, and then what happened?	And then he got into bed.
Hm-hm. Tell me about the part where you said you heard a noise.	We heard a noise and I think it was my aunty and uncle going to bed.
Hm-hm. And does anyone else apart form mum know that Callum had put his dirty spot in your bottom?	No.
Okay, so where were your aunty and uncle when Callum put his dirty spot in your bottom?	I think they were asleep.
Okay, and has Callum ever said anything to you about putting his dirty spot in your bottom?	No.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Oh, okay, yeah.	And then he put something into my bottom.
Yeah, okay, something in your bottom. Then, what happened?	And then it really hurt and I started crying.
Okay. Did anything else happen?	Yes, he pulled it out and put it back in a few times.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Lots of times.	Yes, and then he just stopped.
Okay, is this the only time that he has hit you?	Yes.
Do you remember how long he was hitting you for?	No.
No? Do you know where he got the stick from?	Yeah, he just got it from the garden.
Okay, and who is Mark?	He's my cousin.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
So, it was just you and Callum?	Yeah.
And you told me that it hurt yeah?	Hm-hm.
And I'm really sorry that that has happened. And did you speak to anybody about what has happened?	Yeah.
Yeah? Who did you speak to?	My mum.
You spoke to your mum. Can you remember when you spoke to mum?	Yeah.
Yeah? When was that?	Just when she picked me up in the morning.
Okay, so it was the next day?	Hm-hm.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Did anybody else see the marks?	No.
No? Why not?	I don't know.
Okay, so tell me more about the part where you broke the controller. What happened?	We were just playing Nintendo and I was just standing up so I could go faster and it fell.
Okay.	And then he said I was in trouble.
Okay, so no one else was home? Where was everyone else?	Work.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Nothing, and what were you doing?	Crying.
Okay, and then what happened when he hit you a hundred times?	Then he left.
He did what? Sorry?	He left.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Tell me more about the part where he put his hand in your undies.	He wrapped his hand around my doodle and then I said I didn't like it.
Hm-hm.	And then he laughed and took it out.
Okay, what happens next?	And then I went home.
Hm-hm, and was anyone else with you at uncle George?	No.
Hm-hm, and can you tell me more about the big boys game? Have you played it more than one time?	Yes.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Up to your chin, all right. So tell me about the bit when he tickles your mini. How long does he do this for?	A long time.
A long time. What does it feel like?	A bit yucky.
Okay, all right. When this happens where is your mum?	She is at work.
Hm-hm. I want to find out a bit more about where this happened. Where do you have your bath?	In the bathroom. In the big one.
In the big bathroom, okay. So I know a little bit about your family, but can you tell me a bit more about your family? I know you have got a mum and Stephen. Can you tell me more about your family?	Yes.
Tell me. Who is in your family?	My real dad, but he doesn't live with us and then we have one dog.
Oh okay. So there is mum and Stephen and you have one dog?	Hm-hm.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
It's Lewis. Okay, and do you know what's daddy's surname?	Hm.
Last name.	It's Brown.
Okay, okay, so dad's name is Lewis Brown. Okay, and when did you last see daddy?	About two weeks ago.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Oh, yeah. And what does that mean?	He tells me to go on my hands and knees.
Hm-hm.	And then he put the stick in my bottom.
Oh, okay. Has this happened before?	No.
And then what happened?	I started crying.
Okay. Did it hurt?	Yeah.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Okay, so it's very important that you tell me what has happened in the tunnel. It's very important that you tell me what happened in the tunnel. Are you able to talk with me a bit more about that?	Yeah.
Okay, what happened in the tunnel then Isabelle?	He told me that he has to be the doctor and I have to be the patient.
Okay, and then what happens with that game?	He got a stick and he knocked it on all my bones.
And he what? Sorry.	Knocked it on all my bones.
Knocked it on your bones?	Hm-hm.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Okay, and what's that?	My undies.
Okay, and what's your undies like?	They are white.
They're white undies. Okay, good girl. Okay, that's brilliant. You've given me lots of information there Sophie. Okay, wo where would your clothes be now?	He pulled my undies off.
Okay, and where did they go?	I don't know.
Okay, that's fine. And when you talked to mummy, did you tell mummy that your undies were off?	No.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Now, right at that point where Mark is looking angry, tell me everything that happened from that point till the point where he asked you to lie down on your tummy. Tell me what he said and where you were.	He told me to go to my room and lie down on my tummy.
Okay, and when you laid down on your tummy. Where was that?	On my bed.
On your bed. And what did you think about that?	I was scared.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Okay, did you say anything to the lady?	No.
No? So when she asked if you are okay, you didn't say anything to her?	No.
No? Okay, so what did you do after that?	I went back to mum.
You went back to your mum. Okay, and did you tell your mum what happened?	Yeah.
Yeah? Can you remember what you said to your mum?	Yeah.
Yeah? What did you say?	I told her I was scared going to the toilet.
Okay, and Jessica, do you remember what you were wearing when you went to the toilet?	Hm-hm.

What emotion best fits the child's answers? *

- ◯ Enjoyment
- ◯ Fear
- ◯ Sadness
- ◯ Anger

Interviewer	Child
Hm-hm.	And I told her that I don't want to go back to Janet's house anymore.
Okay. And what happened when you told your mummy you didn't want to go to Janet's house anymore?	She just asked me why.
Hm-hm.	And I just said because Janet's pop he touched me and I didn't like it.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
At lunchtime. And tell me everything that happens when you go into the toilet with Daniel at lunchtime. Right from the beginning when you first go into the toilet. What happens?	Daniel and I, we were playing on the monkey bars and we needed to go to the toilet so we went together.
Okay, and then what happened?	When we got in the toilet, one of the teachers, he came in.
Hm-hm.	And he told us that he wanted to teach us a new game.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm.	So that's what I was doing and he was helping me.
Hm-hm. And can you tell me more about what happened in the cupboard? Because remember, I wasn't there so I don't know what happened.	He said that the game was called show and tell.
Hm-hm.	But when I play normal show and tell, it is not with your privates.
Hm-hm, and so what happened when there was a surprise?	It was just when he showed me his gross willy.
Hm-hm.	And it made me really scared.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm.	And so that's when I started crying a lot and I couldn't stop crying.
Hm-hm. And what happened next?	Then he just told me that we don't have to play anymore.
Hm-hm.	And so he just pulled his pants back up.
Hm-hm.	And I let go of my dress.
Hm-hm. And then what happened?	Then he just came over and he gave me a hug.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
One time. Okay, can you tell me everything about what happened when your neighbour hurt your pee-pee? Start from the beginning.	Yeah.
Okay, so tell me everything about what happened when your neighbour hurt your pee-pee.	I play hide and seek with him.
Hm-hm.	And we play after school.
Okay.	And he comes over to my house.
Okay, your house.	And he came over when his mum was at work.
Okay, so tell me more. Sorry. I'll start again. So what else happened when your neighbour hurt your pee-pee?	I like hiding first because I'm really good at hiding.
So can you tell me that again? You hide?	First. Because I'm really good at hiding.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Yeah, what makes you feel unsafe?	When Sam makes us go into the toilet.
Okay, how does he make you?	I don't know.
Don't know. When he asks you to go into the toilet, does he say anything to you? What does he say?	He said that we have to go in or he'll bully us at lunchtime.

Interviewer	Child
Hm-hm.	And he was angry because his dinner was cold.
What happened then?	Then he just was throwing his spaghetti at the wall and the plate smashed everywhere.
The plate smashed everywhere?	Yeah, and it made this big red mark on the wall.
Hm, what else happened?	Then that was when he was just being nasty and saying mean things.
Hm-hm.	And that was the part where he hurt mum.
Tell me more about the part where he has hurt your mum?	He just hurt her and she fell down on the ground.
Hm-hm.	And she wasn't moving.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Your cousin.	He's 12 and he has blond hair and he's my only cousin.
Okay. What sort of things do you like to do with your cousin?	Sometimes I stay at his house.
Okay.	And sometimes we play on the computer.
Okay, tell me all about staying at your cousin's house.	He has a single bed and I stay on the blow-up mattress.
Okay, and you stay on the blow-up mattress. Okay, tell me about the last time you stayed at your cousin's house.	He woke me up in the middle of the night.
Sorry, he woke you up in the middle of the night?	Yeah.
Yeah, okay, and what happened?	And then pulled my pants down.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Hm-hm.	And he was angry because his dinner was cold.
What happened then?	Then he just was throwing his spaghetti at the wall and the plate smashed everywhere.
The plate smashed everywhere?	Yeah, and it made this big red mark on the wall.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Interviewer	Child
Okay. Is there anything else you want to tell me?	No.
No? Are you feeling safe at school?	Yes.
Yes? What makes you feel safe?	My friends are really nice.
Friends are nice.	The teachers were nice.
Teachers are nice. Okay, is there anything that makes you feel unsafe?	Yes.
Yeah? What makes you feel unsafe?	When Sam makes us go into the toilet.
Okay, how does he make you?	I don't know.

What emotion best fits the child's answers? *

- Enjoyment
- Fear
- Sadness
- Anger

Send

Responsible for the form: **myrthel@uio.no**.

Appendix B

Appendix B - user studies results

B.1 User Study 1

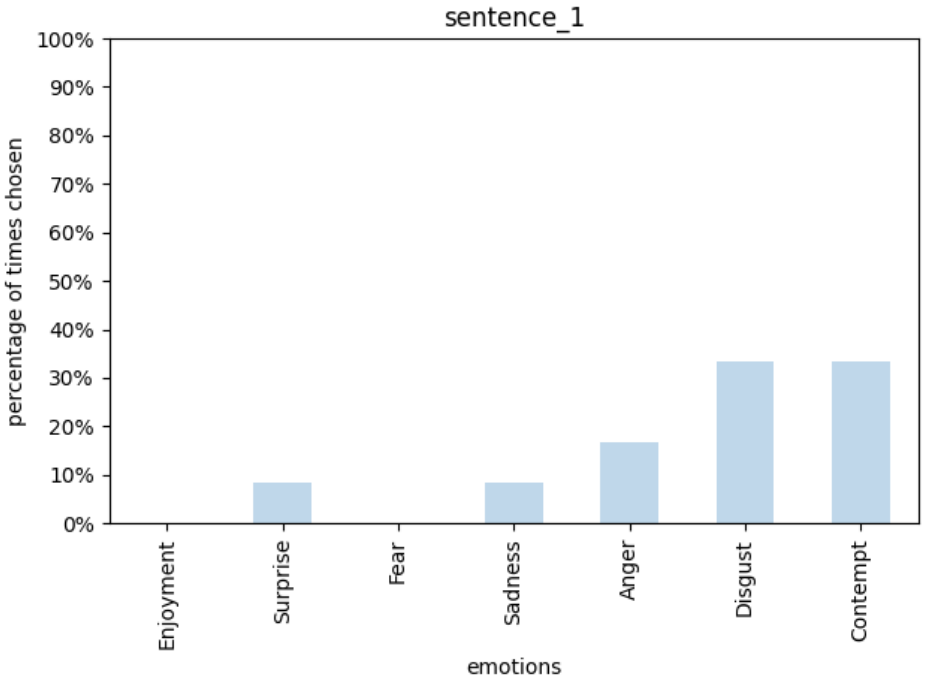

Figure B.1: Results for sentence_1 in user study 1.

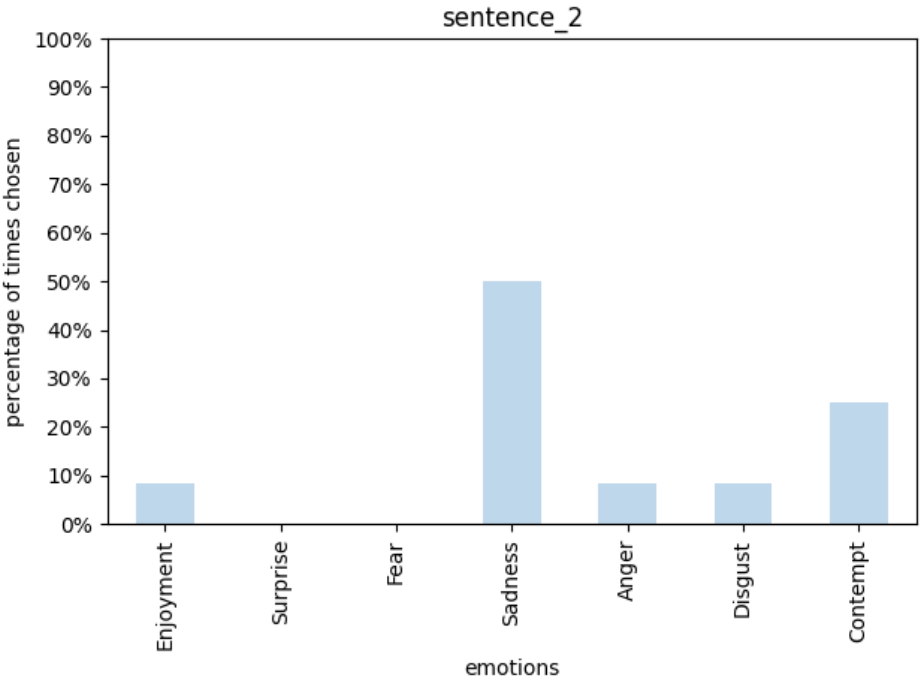

Figure B.2: Results for sentence_2 in user study 1.

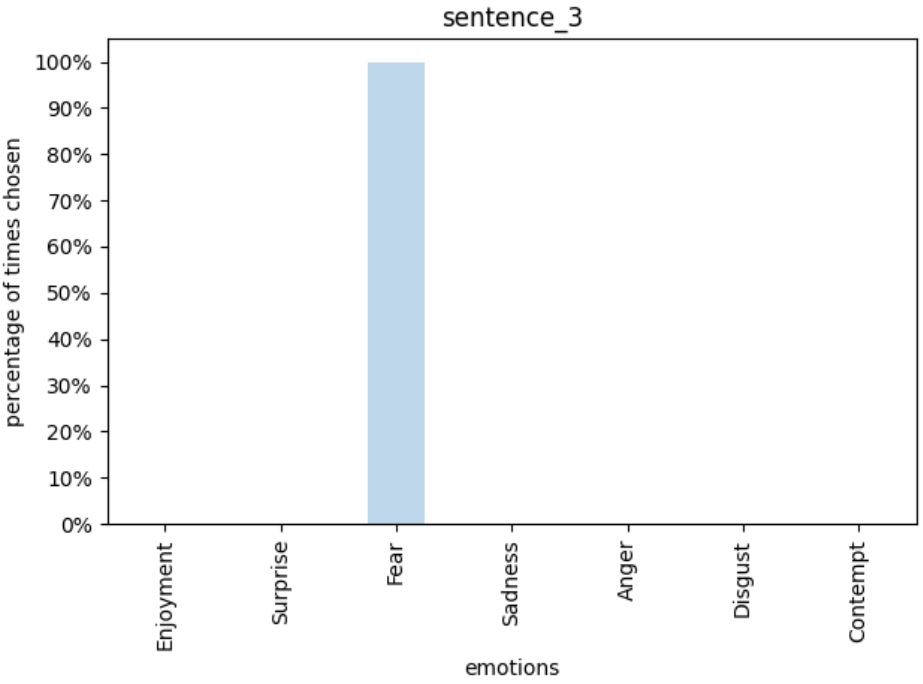

Figure B.3: Results for sentence_3 in user study 1.

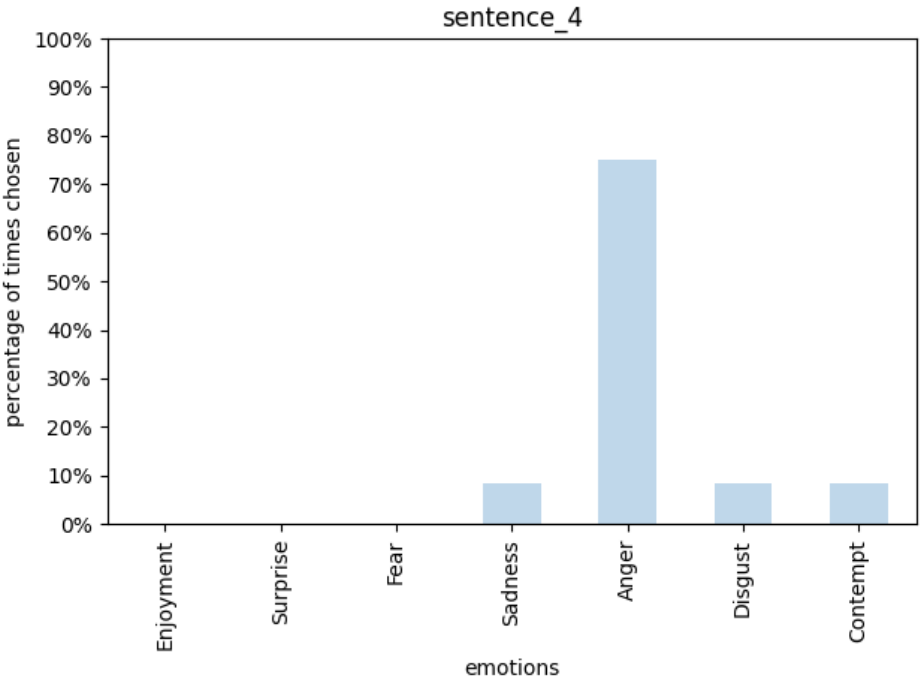

Figure B.4: Results for sentence_4 in user study 1.

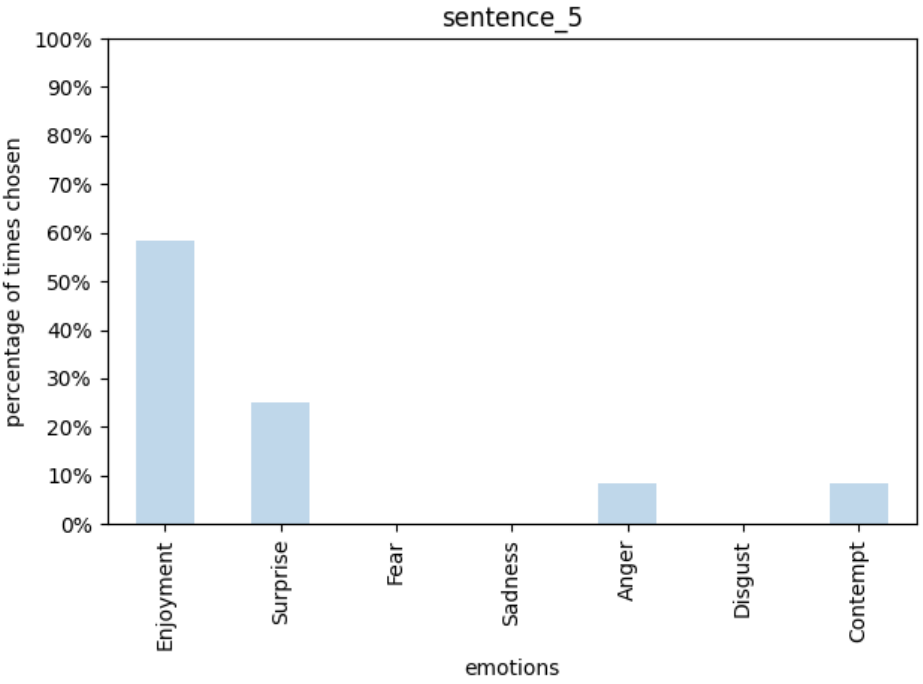

Figure B.5: Results for sentence_5 in user study 1.

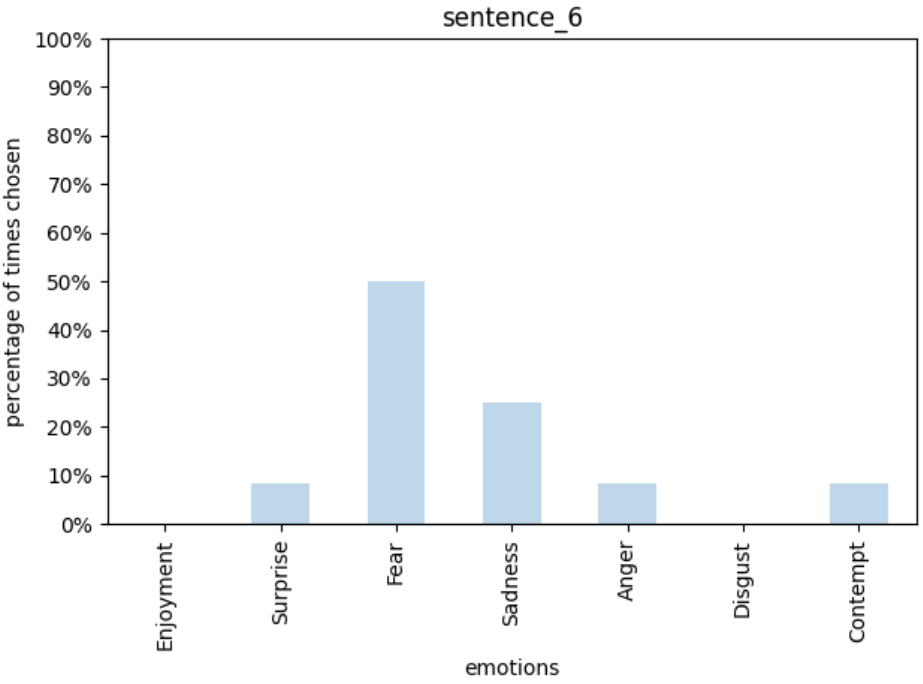

Figure B.6: Results for sentence_6 in user study 1.

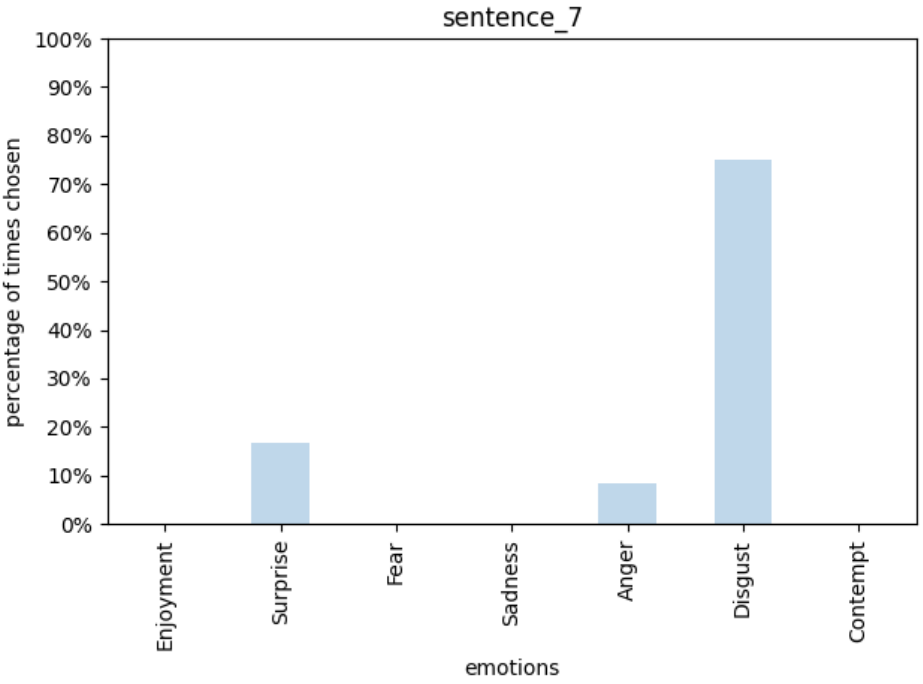

Figure B.7: Results for sentence_7 in user study 1.

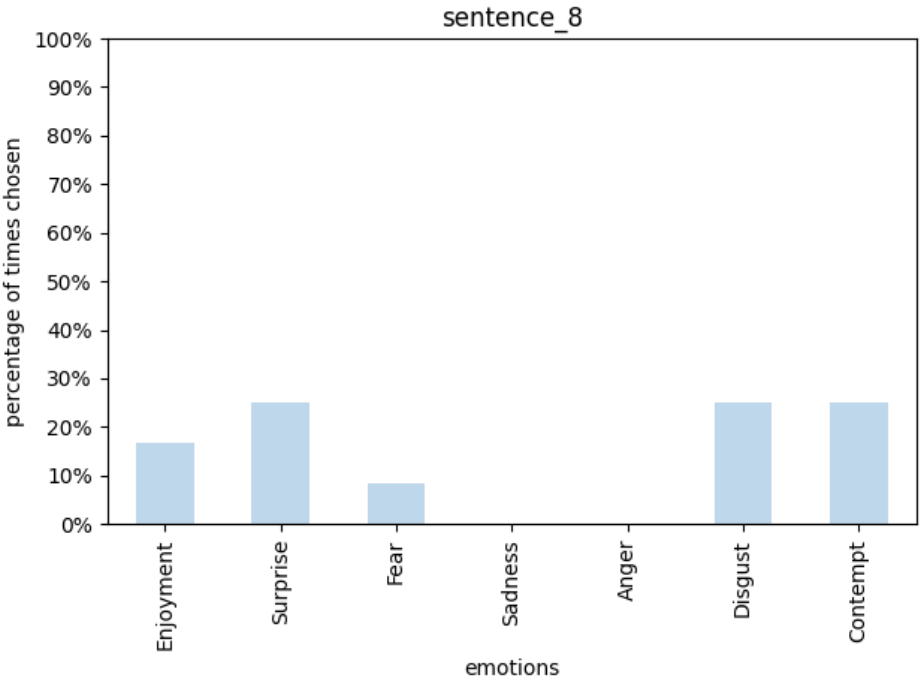

Figure B.8: Results for sentence_8 in user study 1.

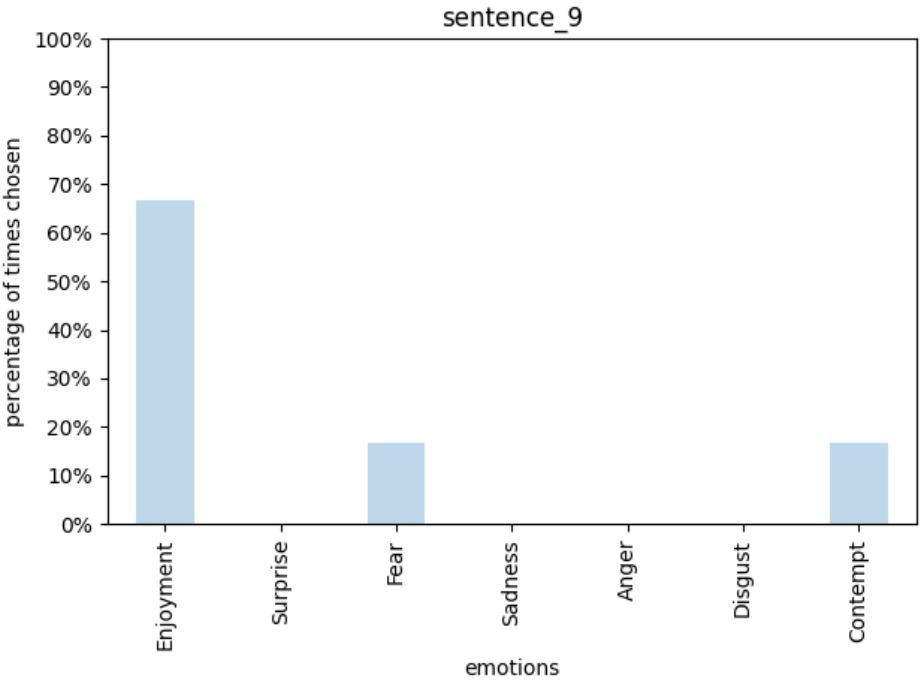

Figure B.9: Results for sentence_9 in user study 1.

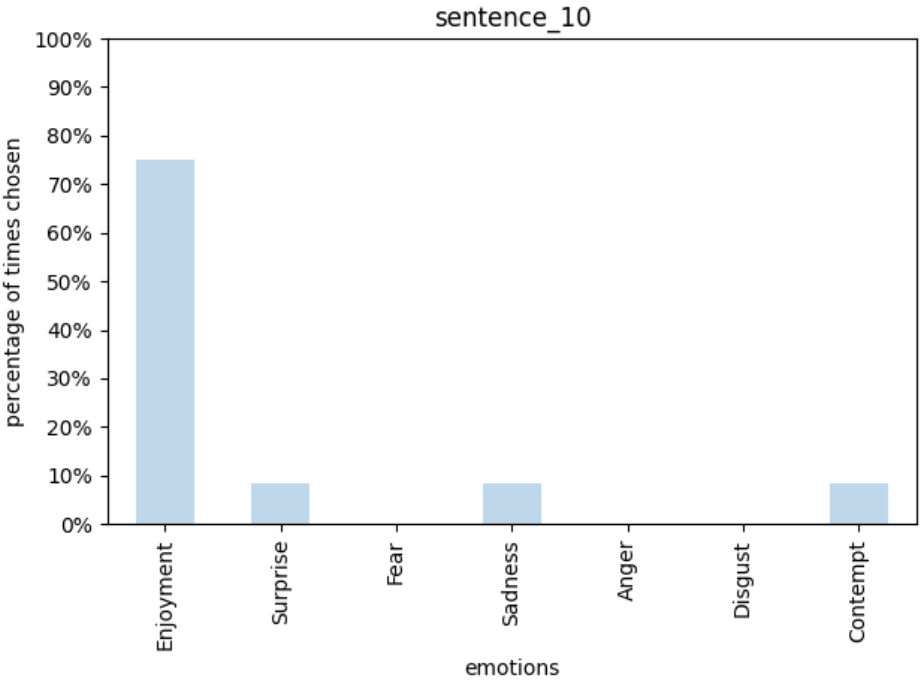

Figure B.10: Results for sentence_10 in user study 1.

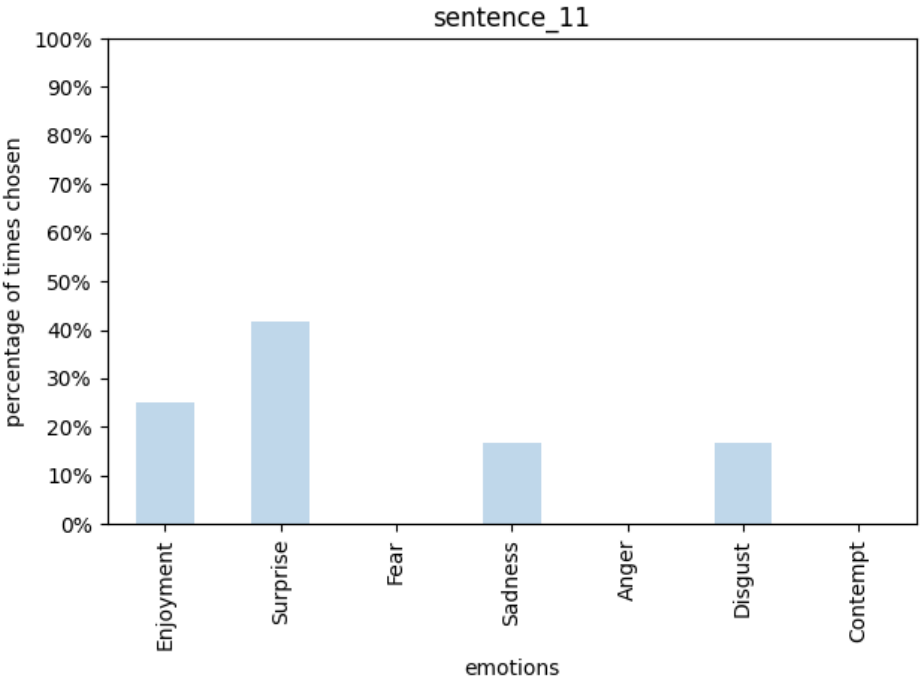

Figure B.11: Results for sentence_11 in user study 1.

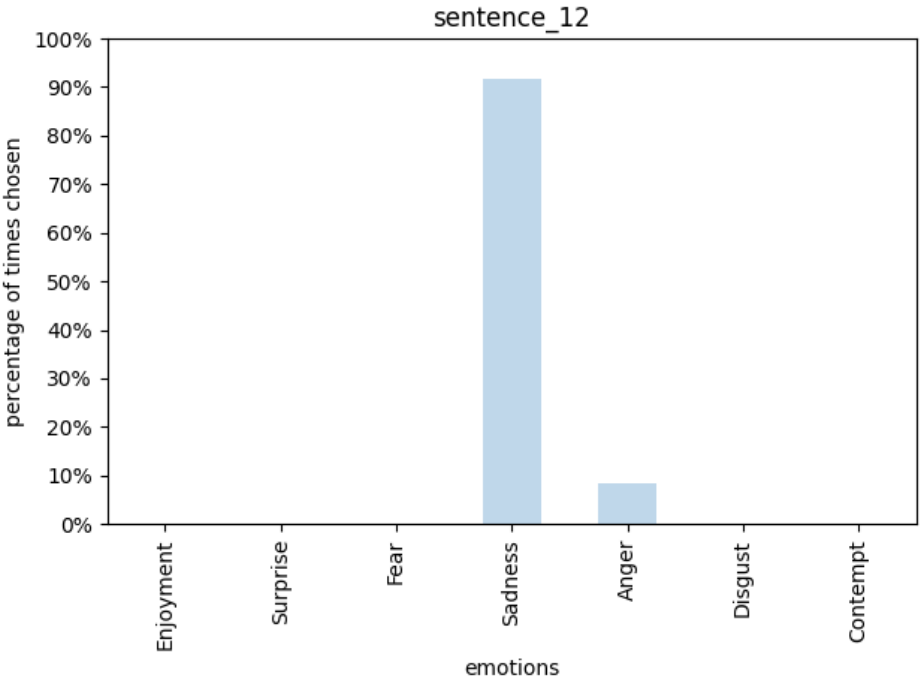

Figure B.12: Results for sentence_12 in user study 1.

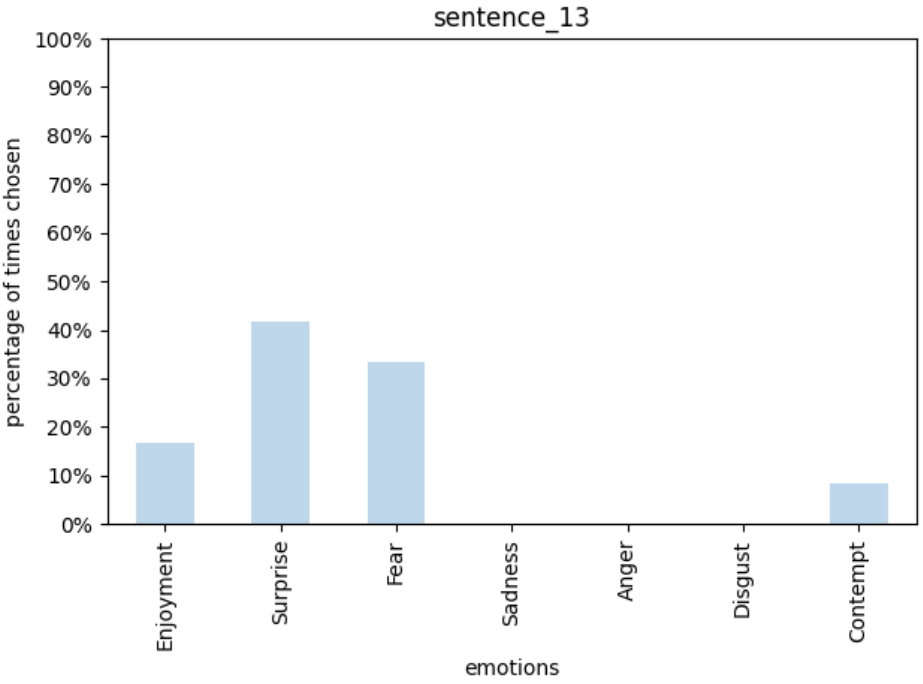

Figure B.13: Results for sentence_13 in user study 1.

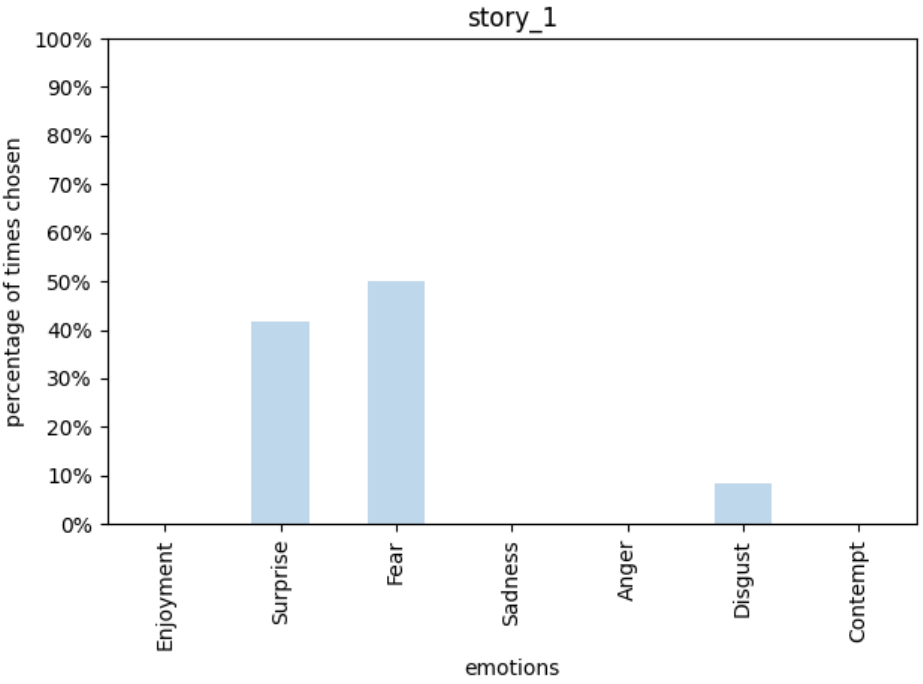

Figure B.14: Results for story_1 in user study 1.

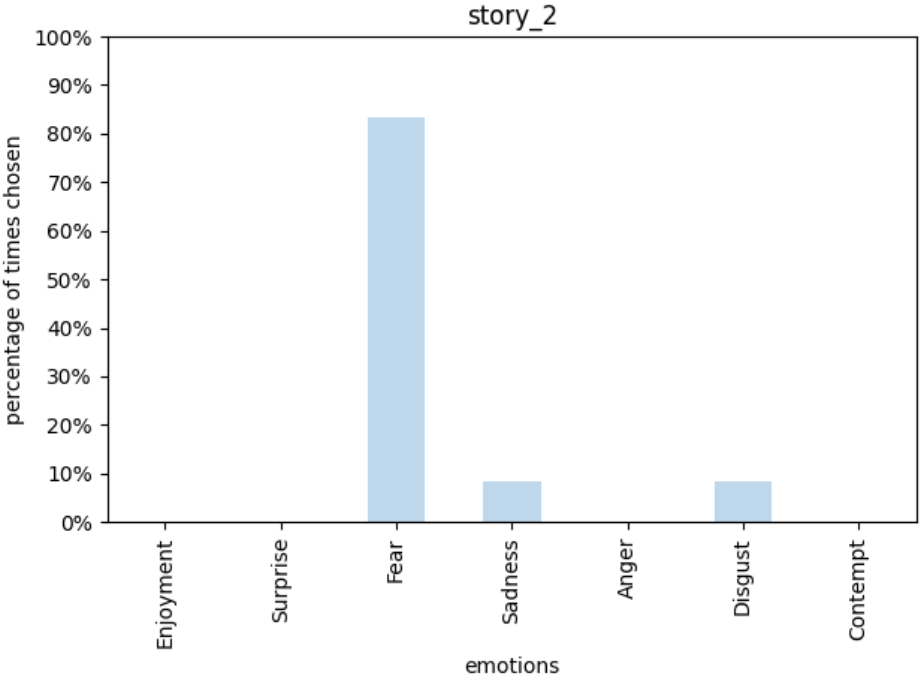

Figure B.15: Results for story_2 in user study 1.

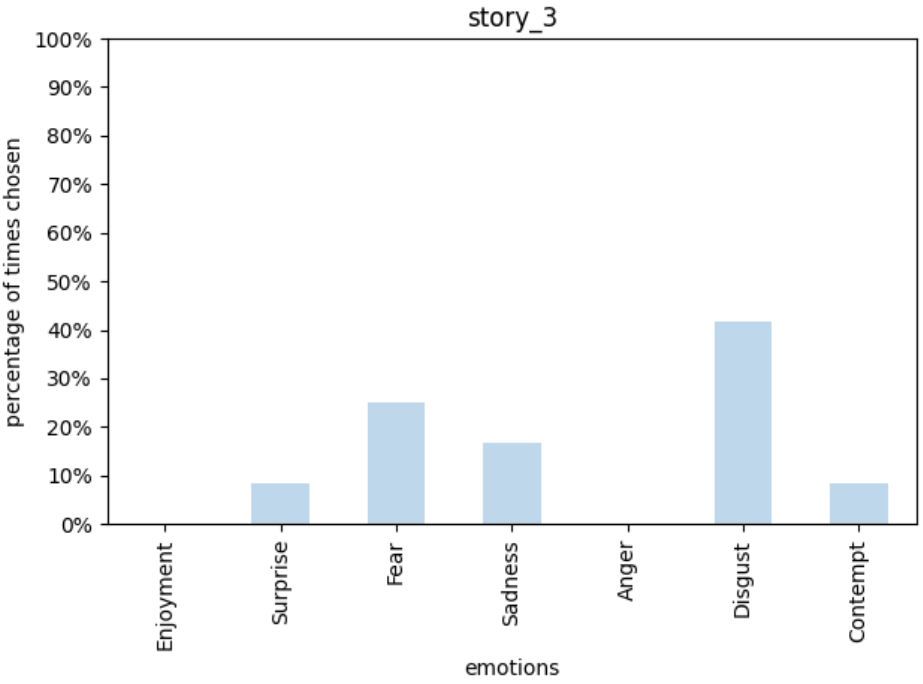

Figure B.16: Results for story_3 in user study 1.

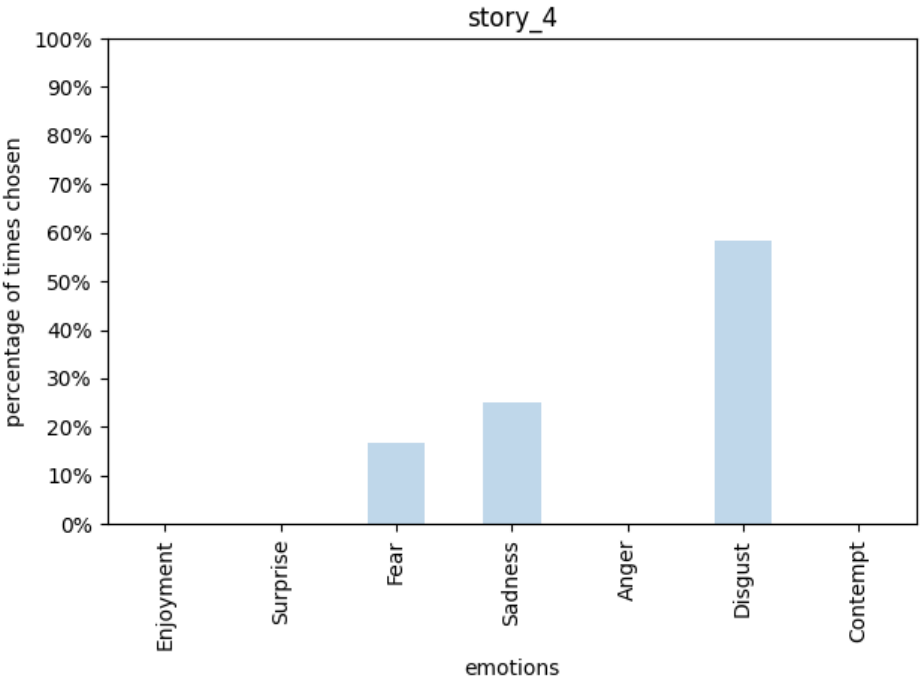

Figure B.17: Results for story_4 in user study 1.

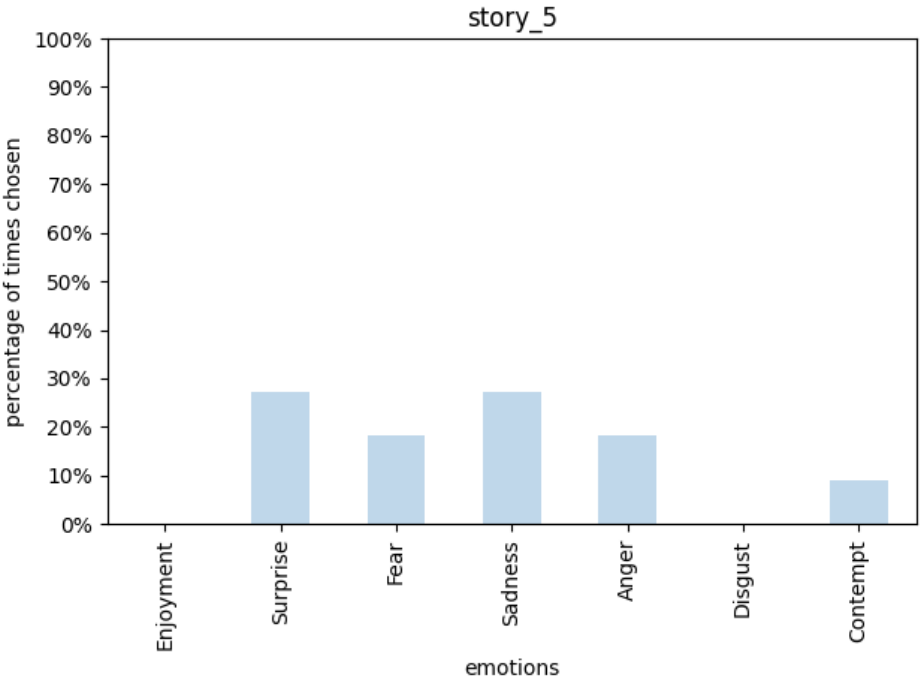

Figure B.18: Results for story_5 in user study 1.

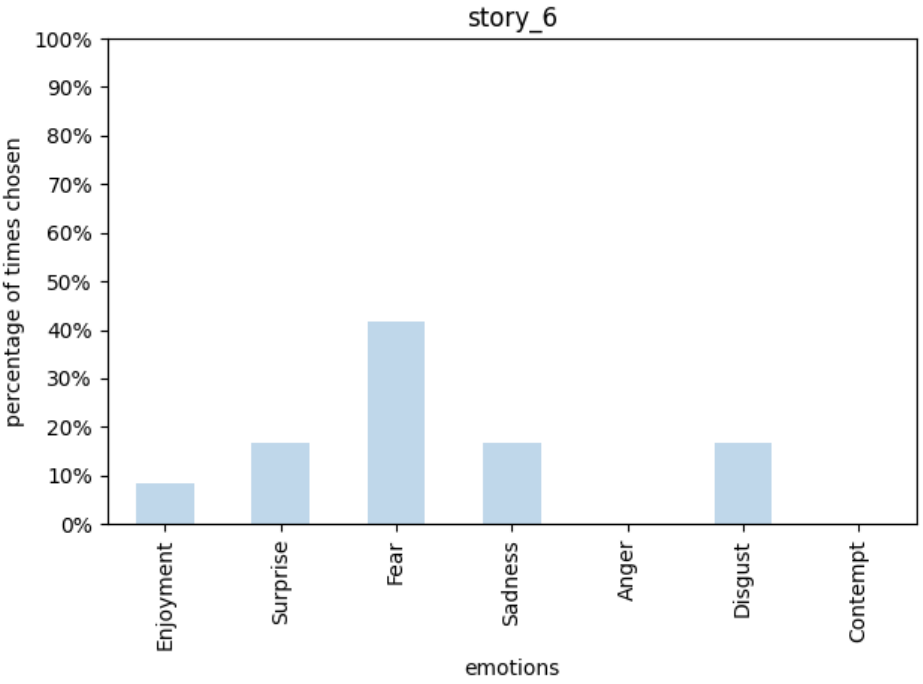

Figure B.19: Results for story_6 in user study 1.

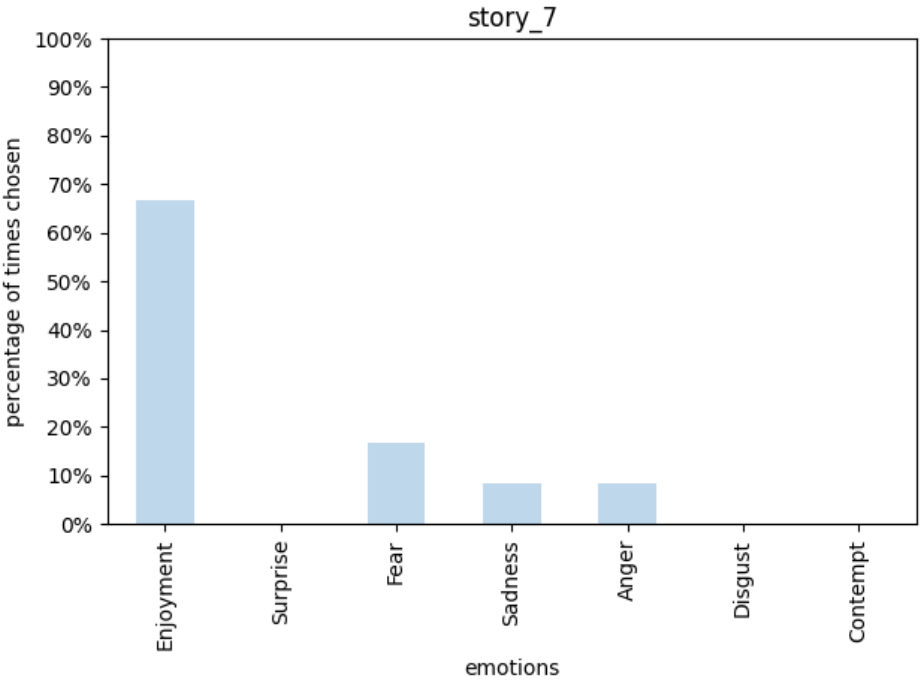

Figure B.20: Results for story_7 in user study 1.

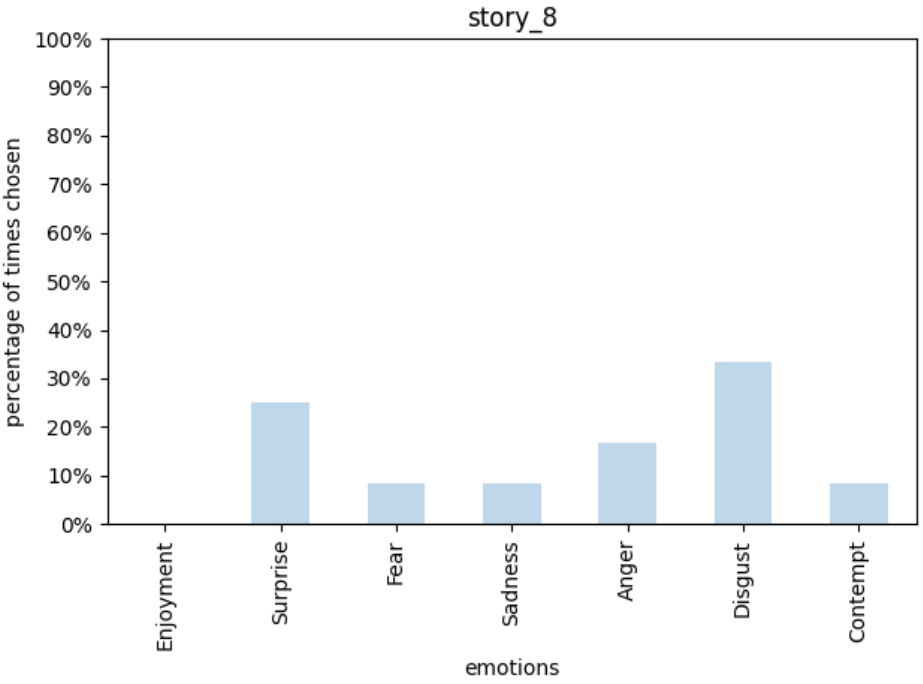

Figure B.21: Results for story_8 in user study 1.

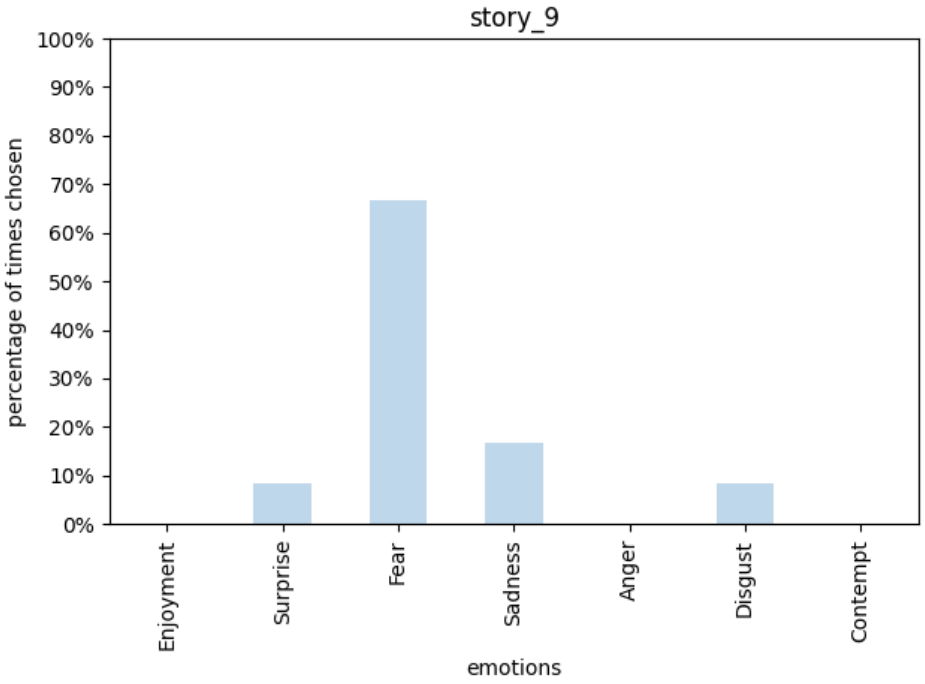

Figure B.22: Results for story_9 in user study 1.

B.2 User Study 2

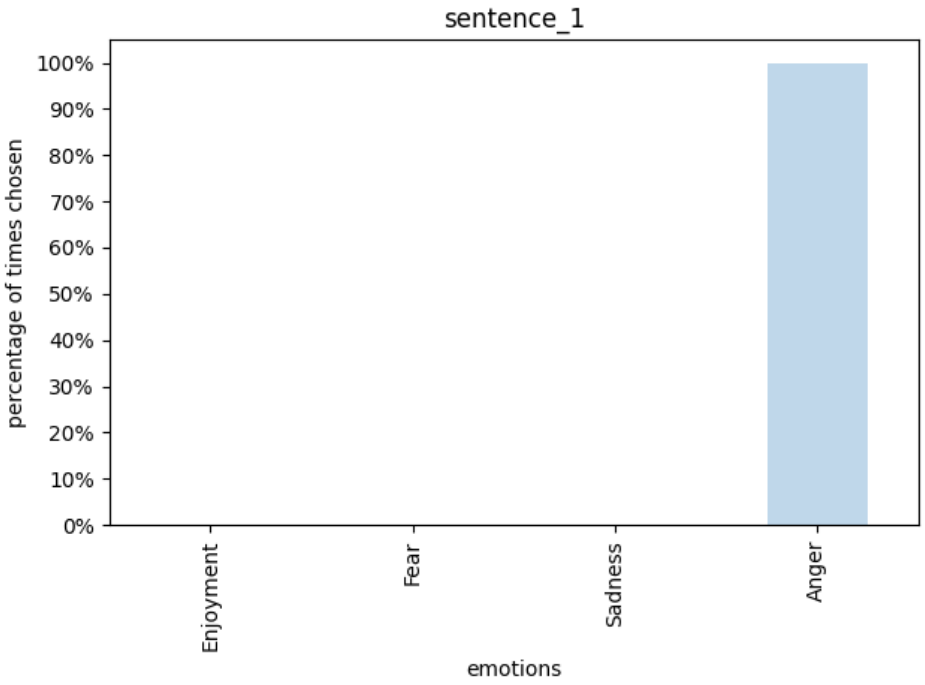

Figure B.23: Results for sentence_1 in user study 2.

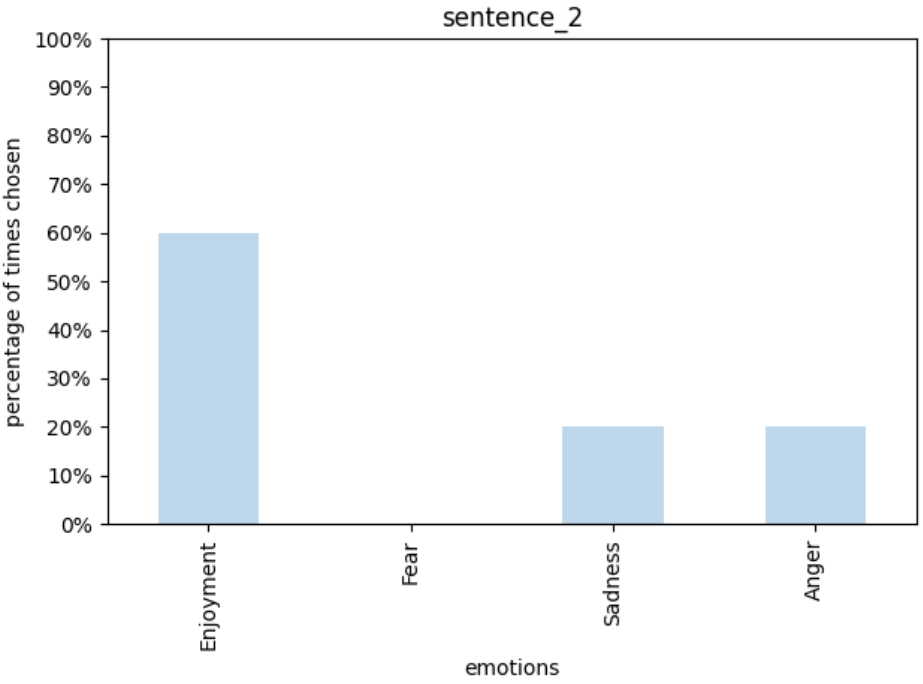

Figure B.24: Results for sentence_2 in user study 2.

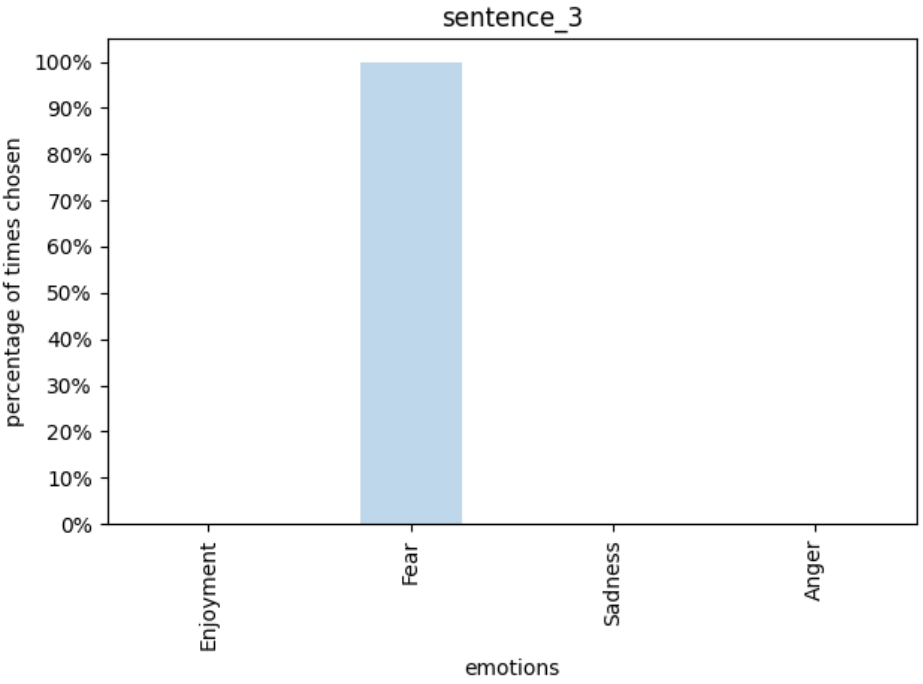

Figure B.25: Results for sentence_3 in user study 2.

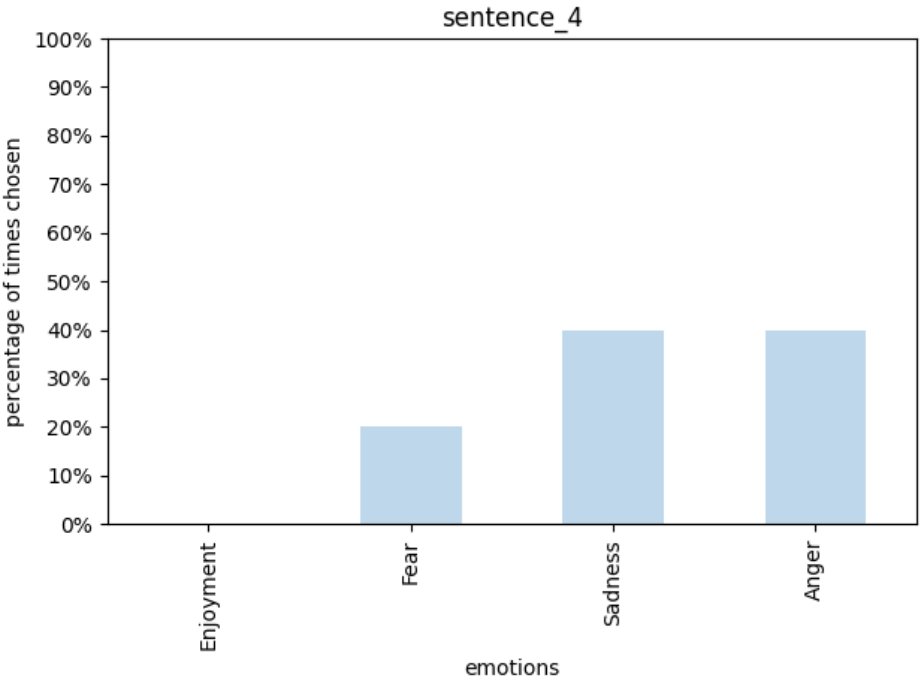

Figure B.26: Results for sentence_4 in user study 2.

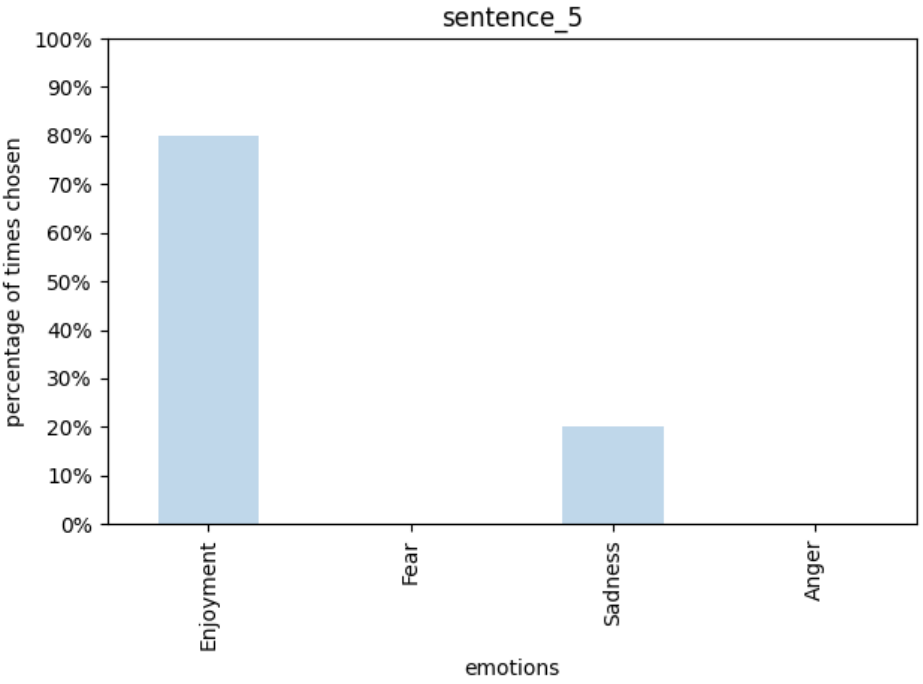

Figure B.27: Results for sentence_5 in user study 2.

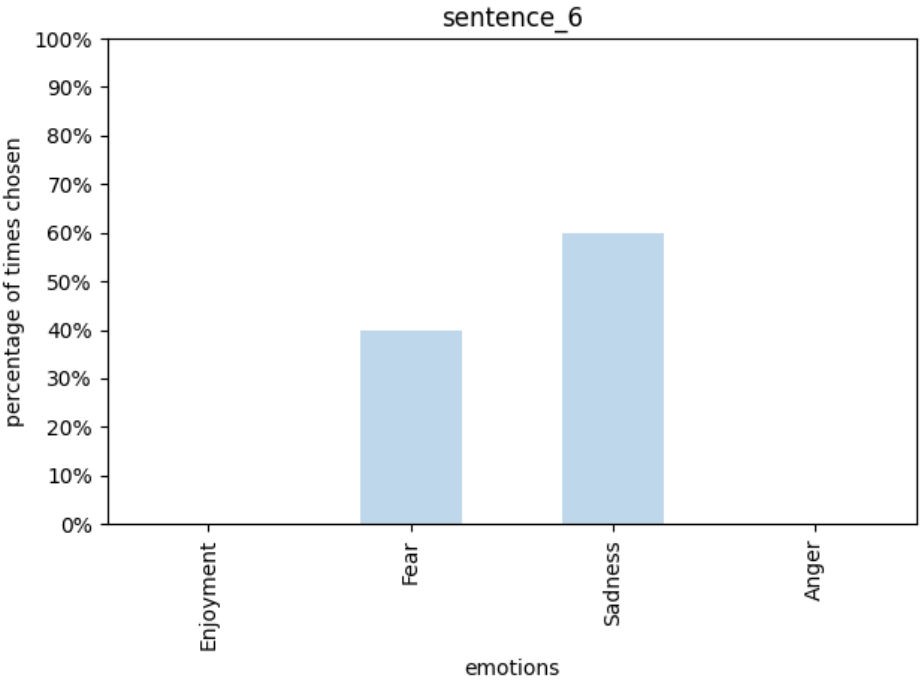

Figure B.28: Results for sentence_6 in user study 2.

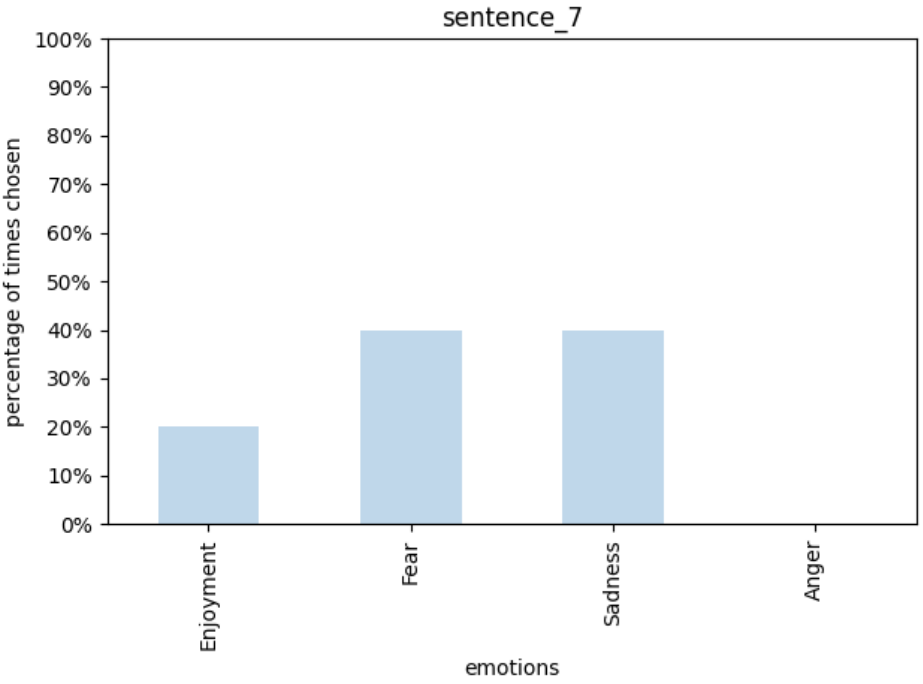

Figure B.29: Results for sentence_7 in user study 2.

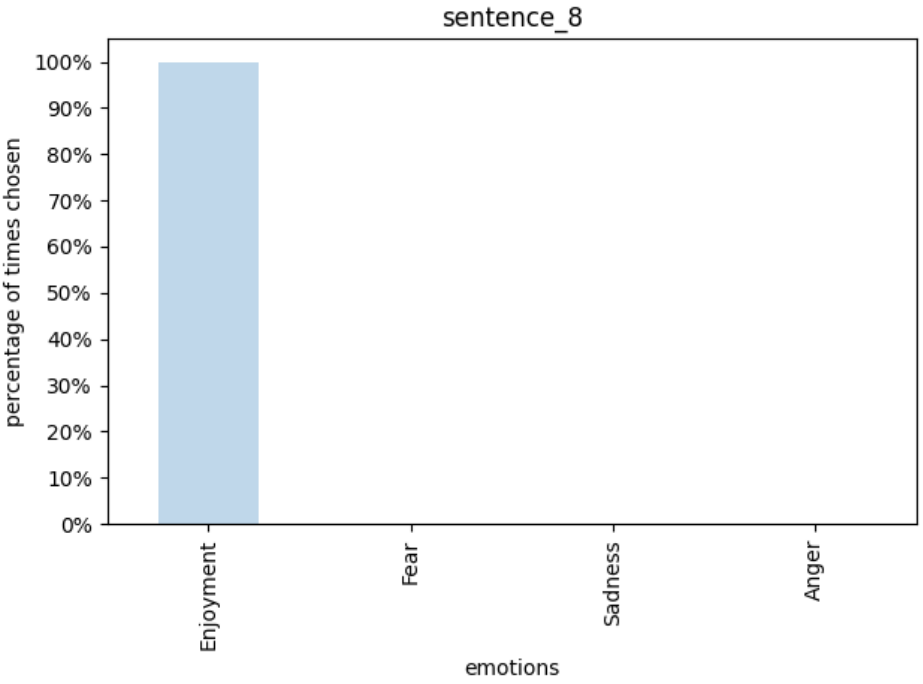

Figure B.30: Results for sentence_8 in user study 2.

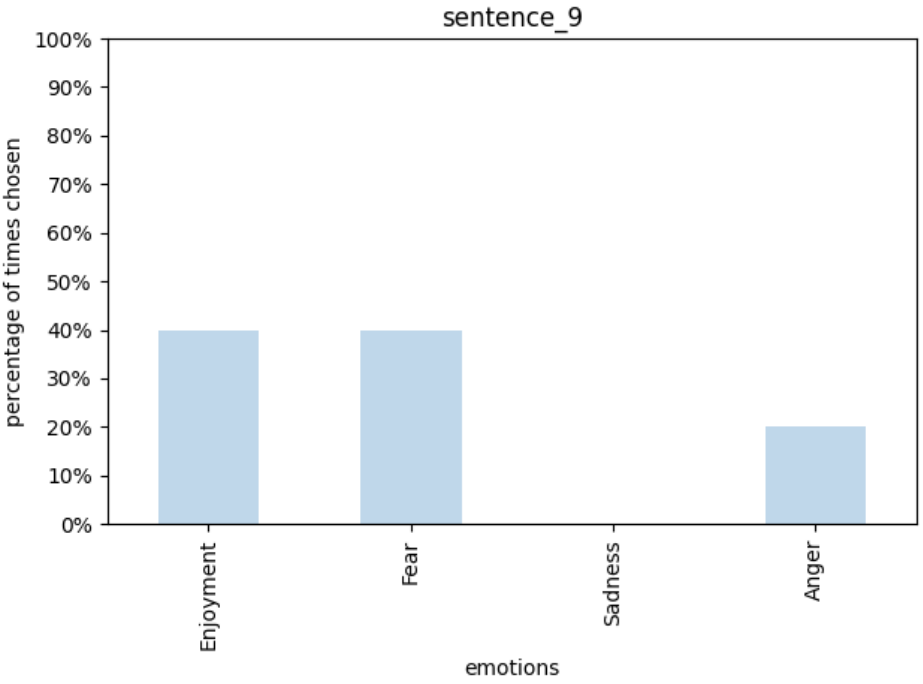

Figure B.31: Results for sentence_9 in user study 2.

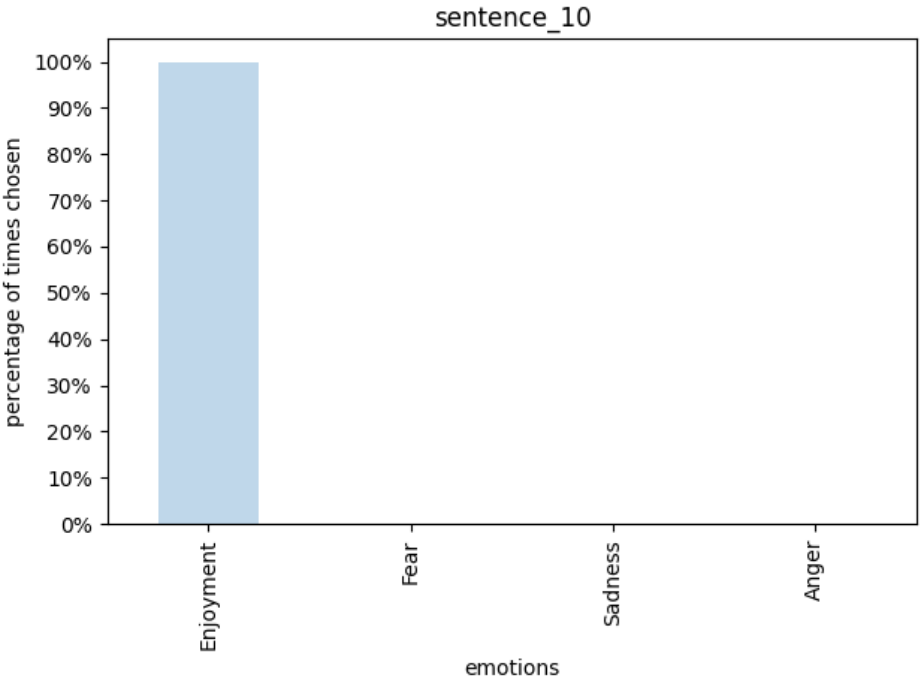

Figure B.32: Results for sentence_10 in user study 2.

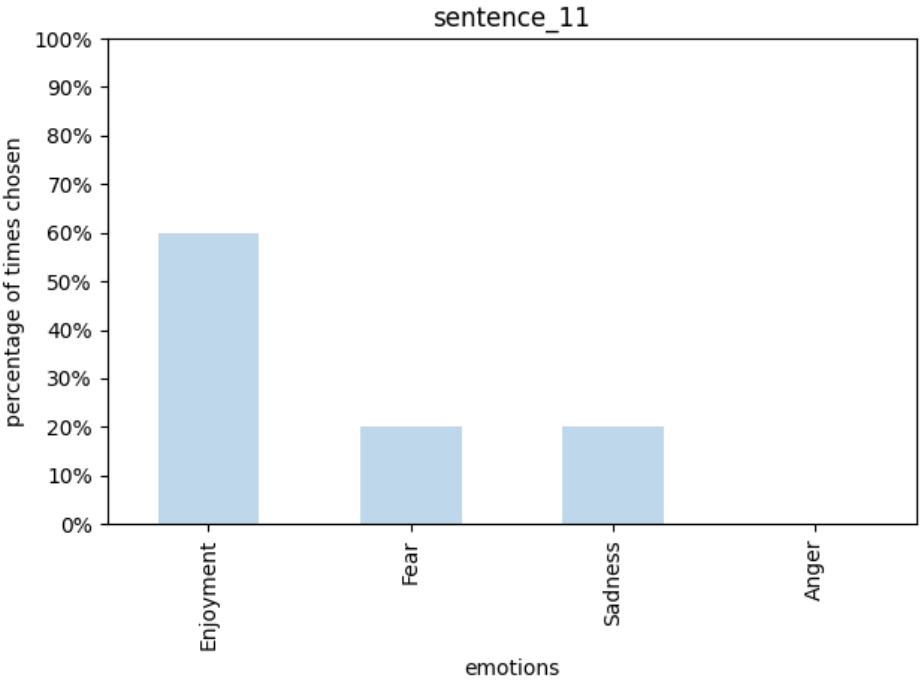

Figure B.33: Results for sentence_11 in user study 2.

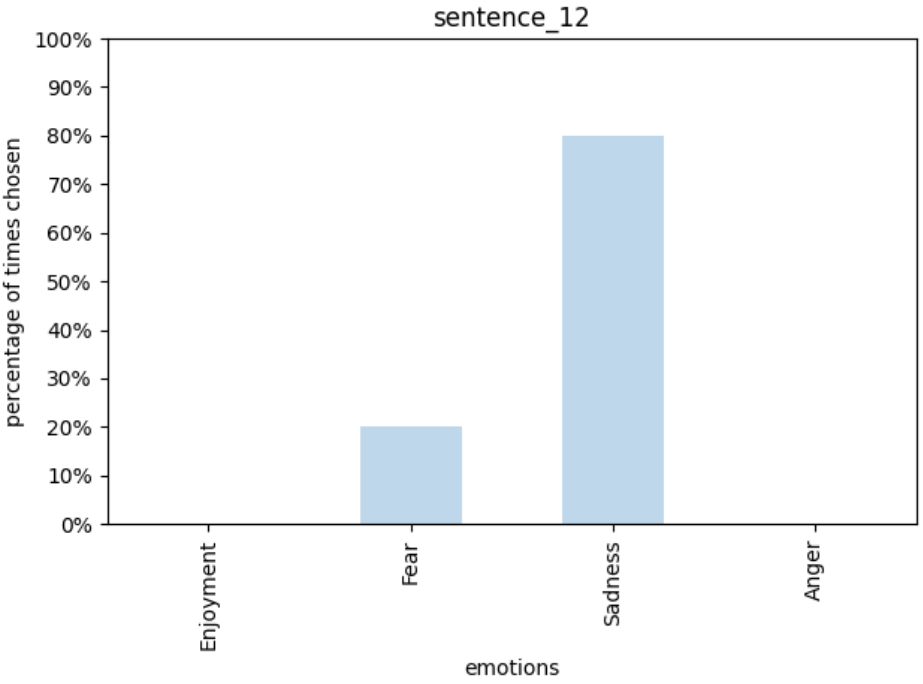

Figure B.34: Results for sentence_12 in user study 2.

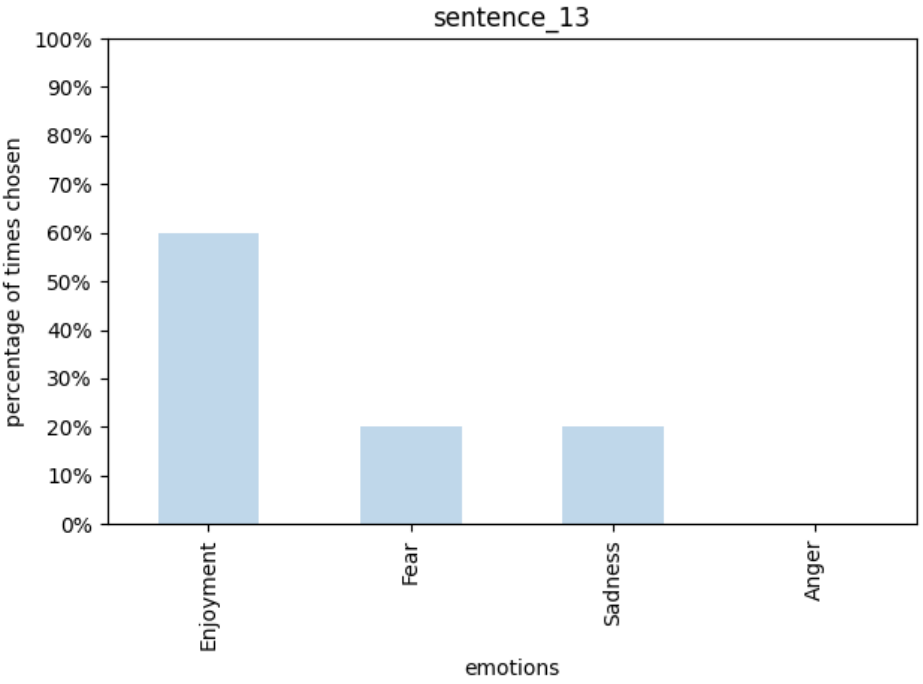

Figure B.35: Results for sentence_13 in user study 2.

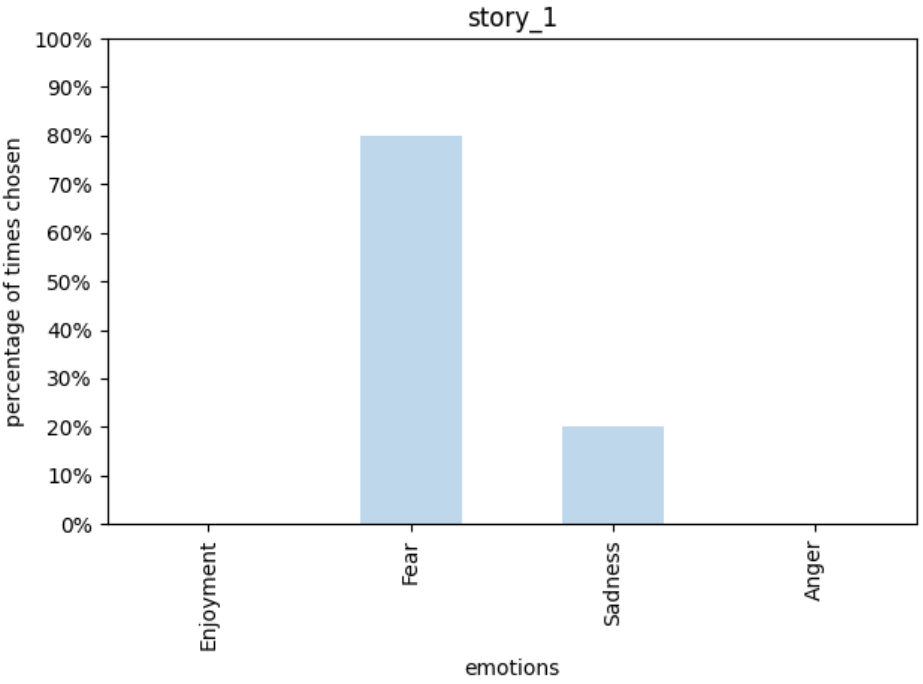

Figure B.36: Results for story_1 in user study 2.

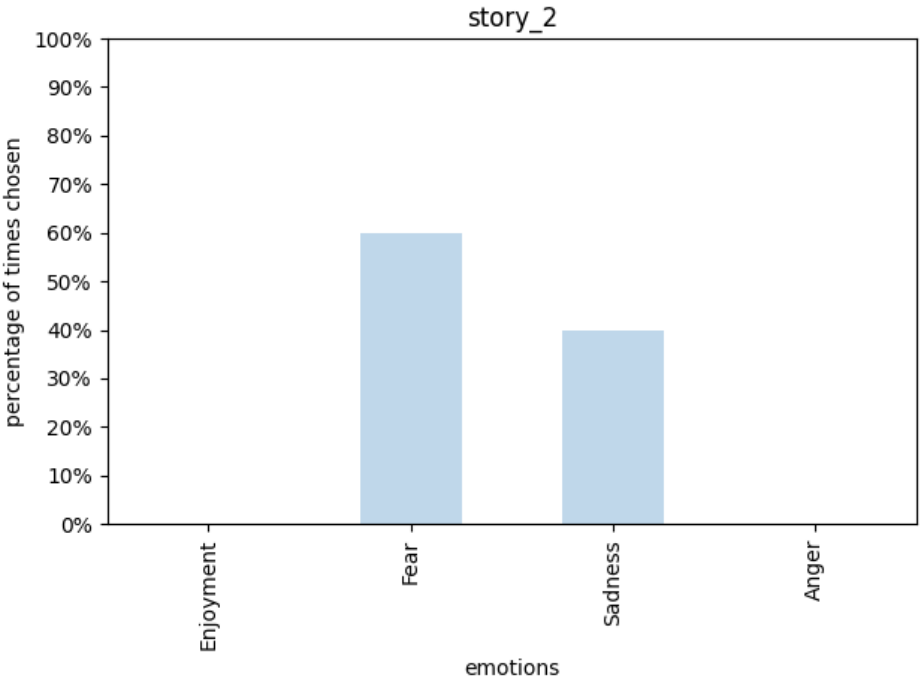

Figure B.37: Results for story_2 in user study 2.

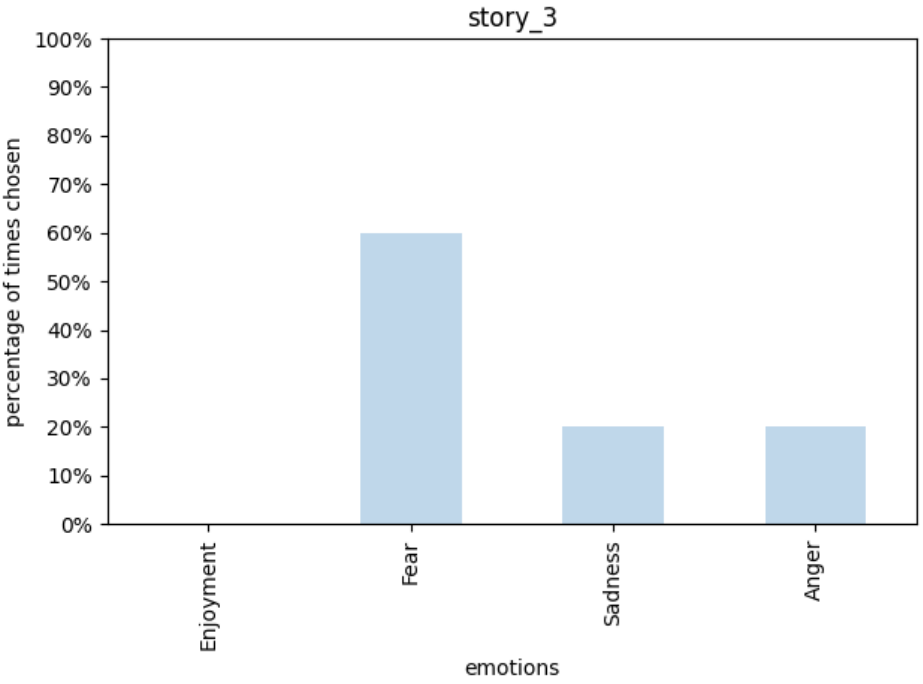

Figure B.38: Results for story_3 in user study 2.

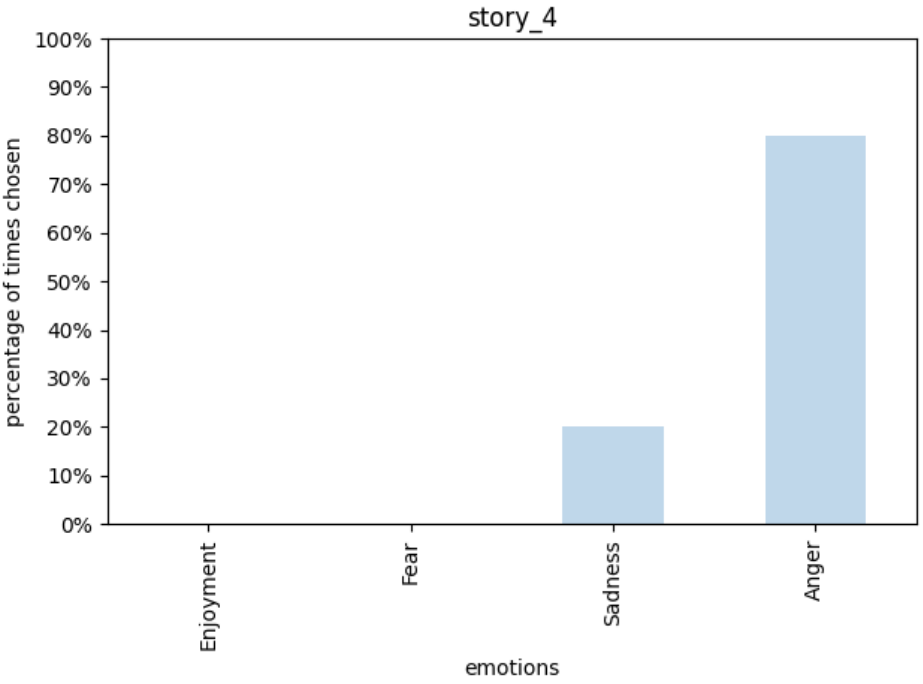

Figure B.39: Results for story_4 in user study 2.

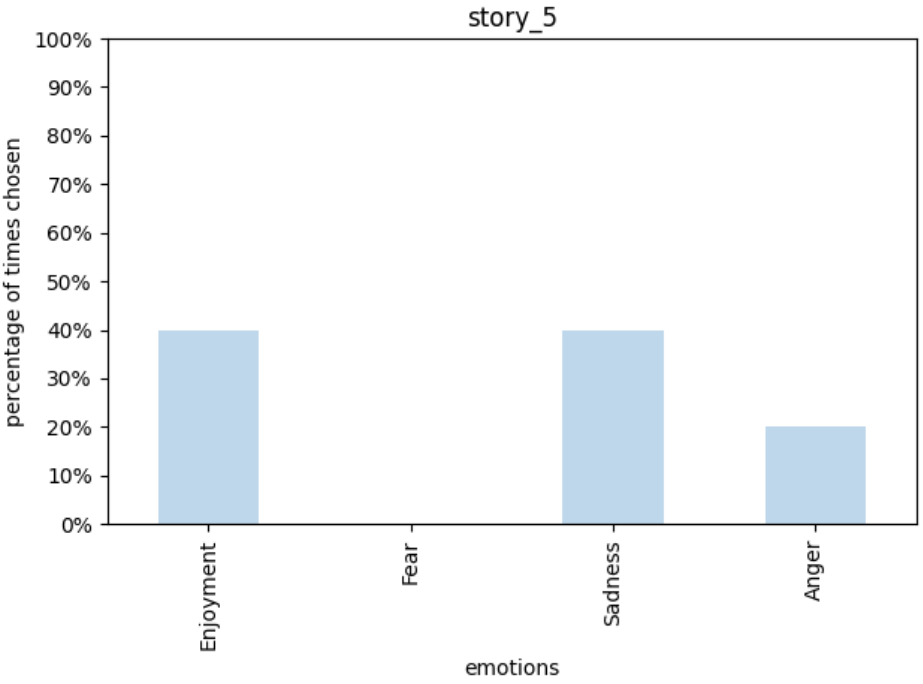

Figure B.40: Results for story_5 in user study 2.

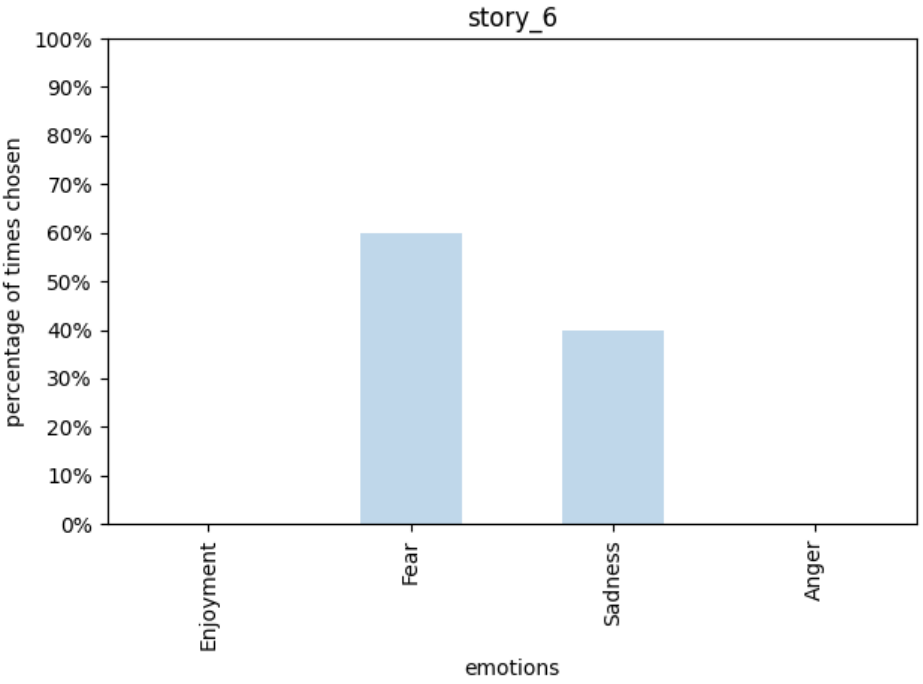

Figure B.41: Results for story_6 in user study 2.

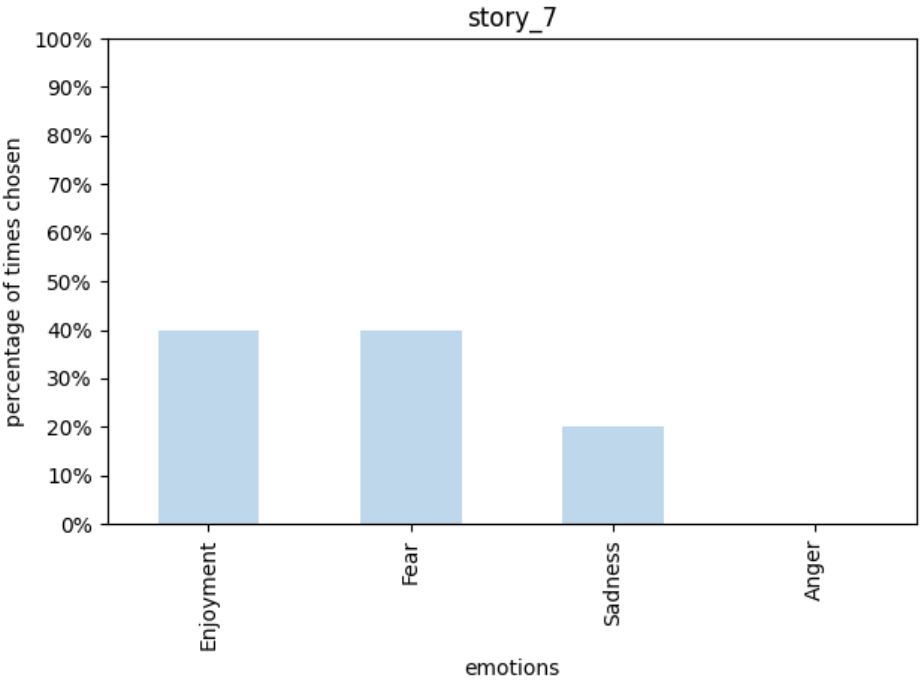

Figure B.42: Results for story_7 in user study 2.

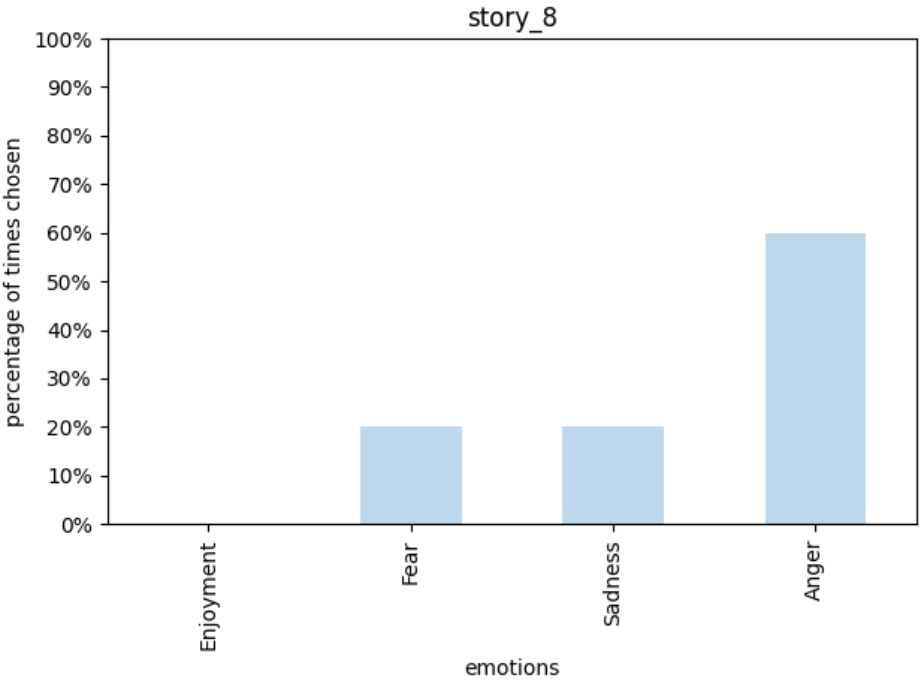

Figure B.43: Results for story_8 in user study 2.

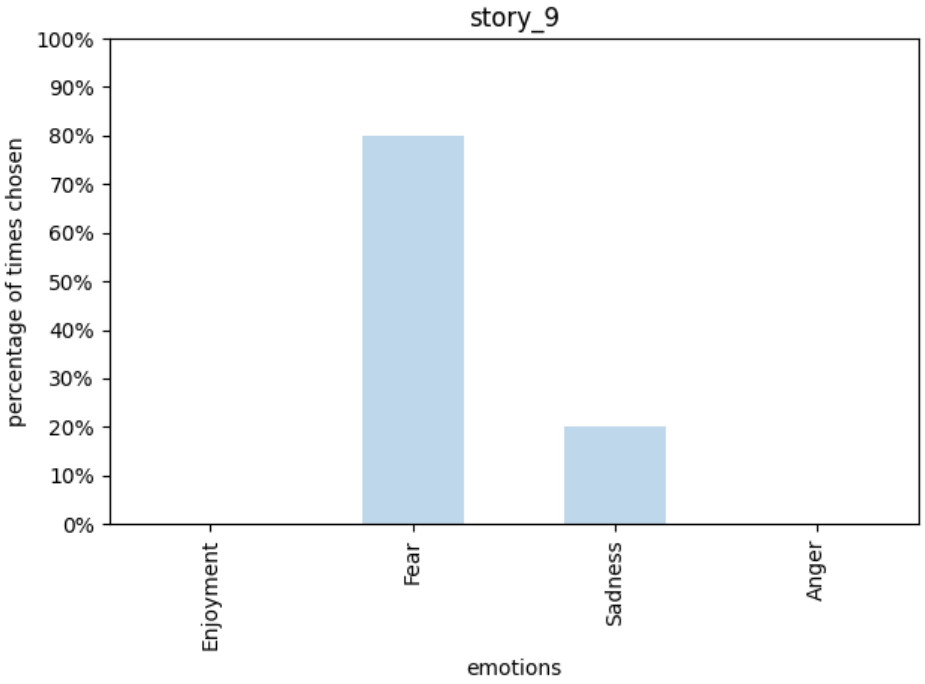

Figure B.44: Results for story_9 in user study 2.

B.3 User Study 3

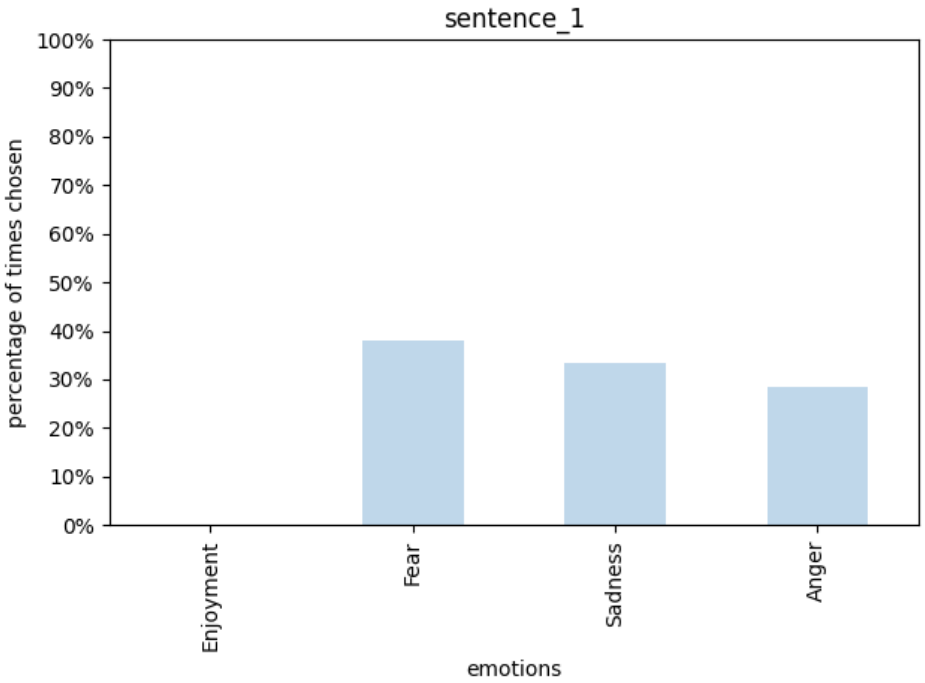

Figure B.45: Results for sentence_1 in user study 3.

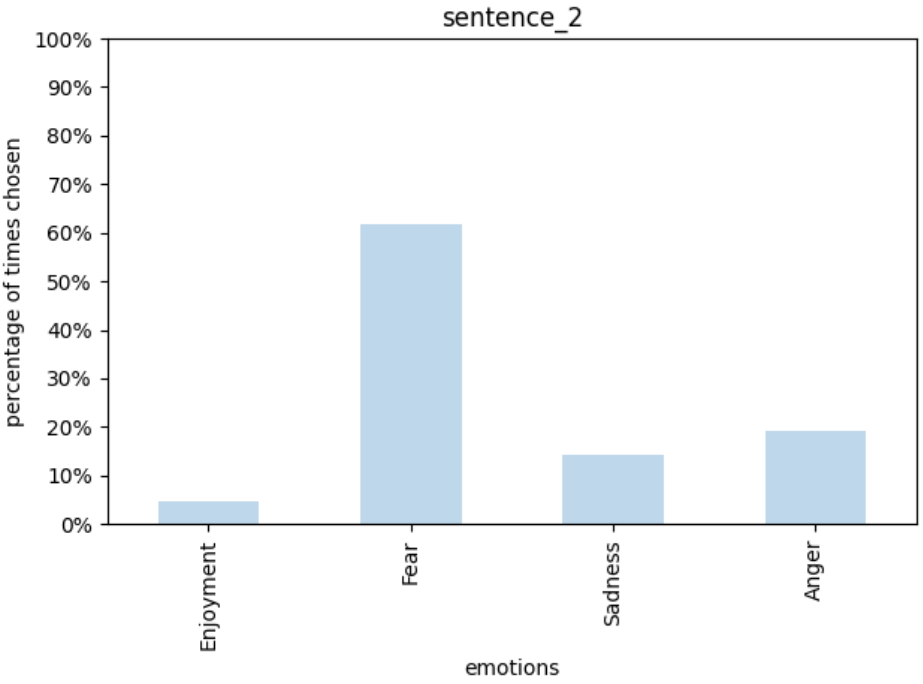

Figure B.46: Results for sentence_2 in user study 3.

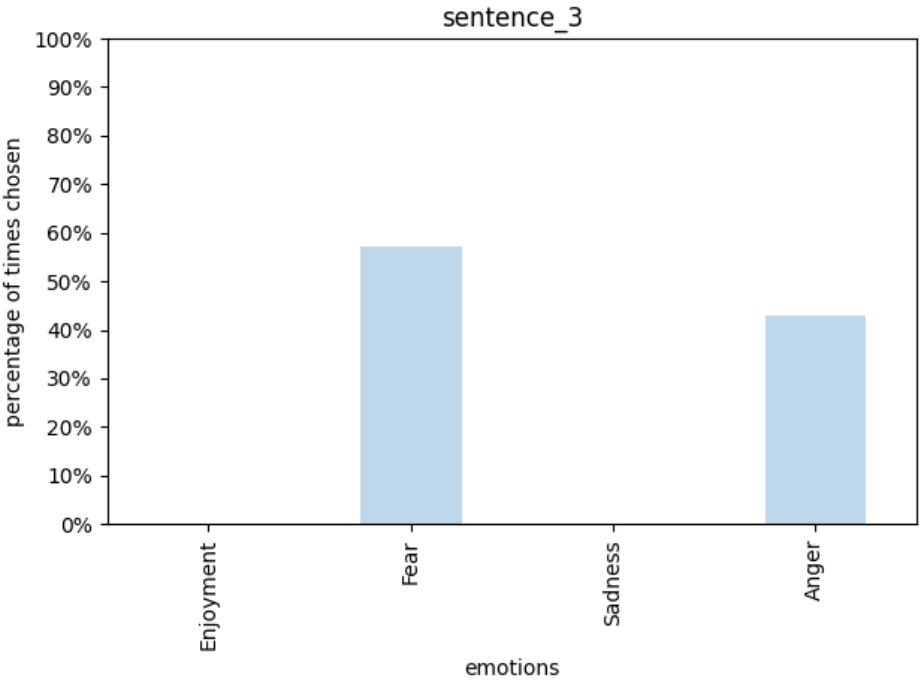

Figure B.47: Results for sentence_3 in user study 3.

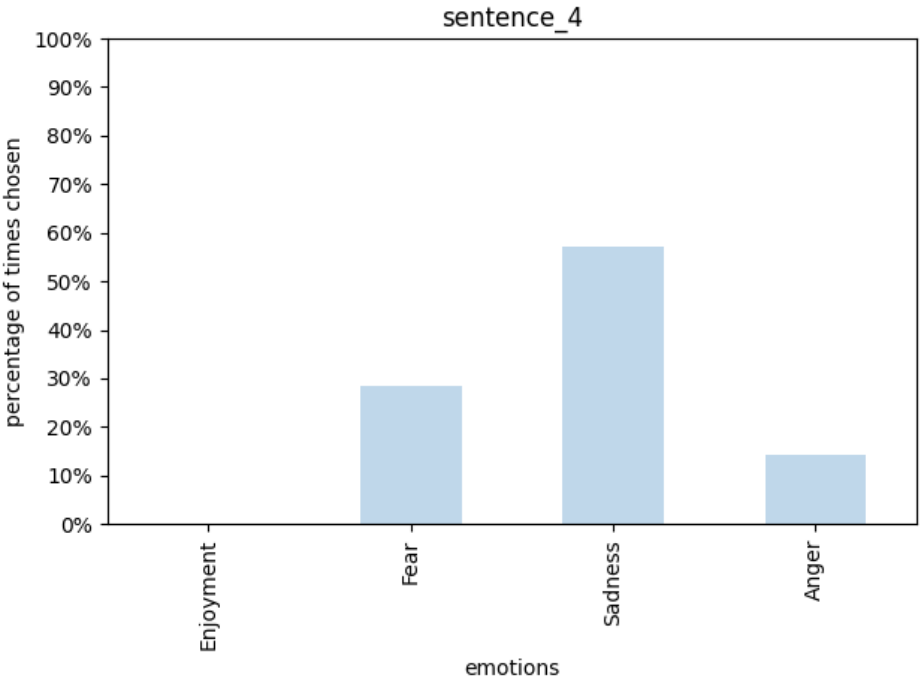

Figure B.48: Results for sentence_4 in user study 3.

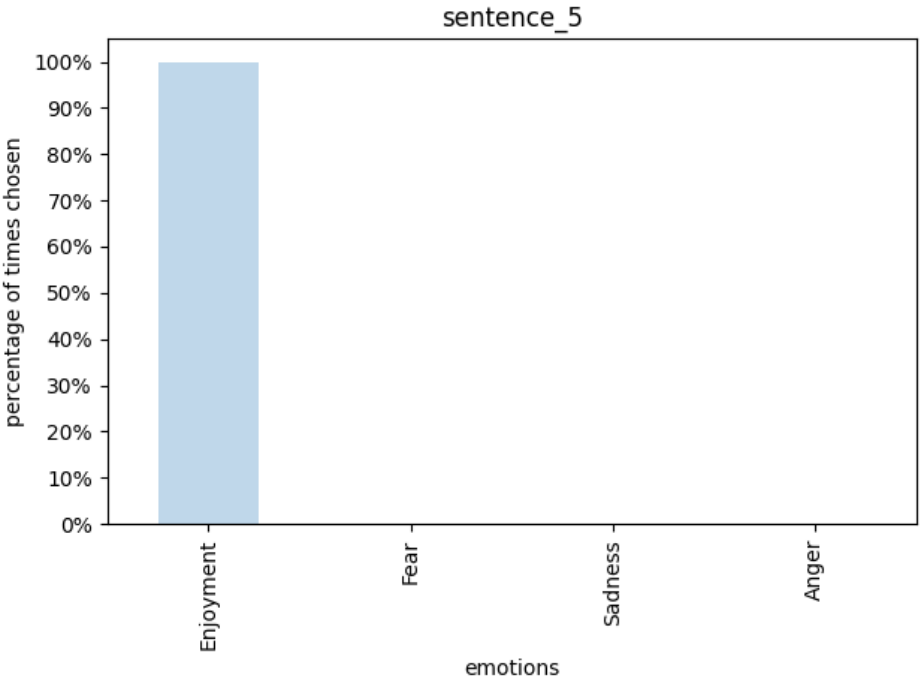

Figure B.49: Results for sentence_5 in user study 3.

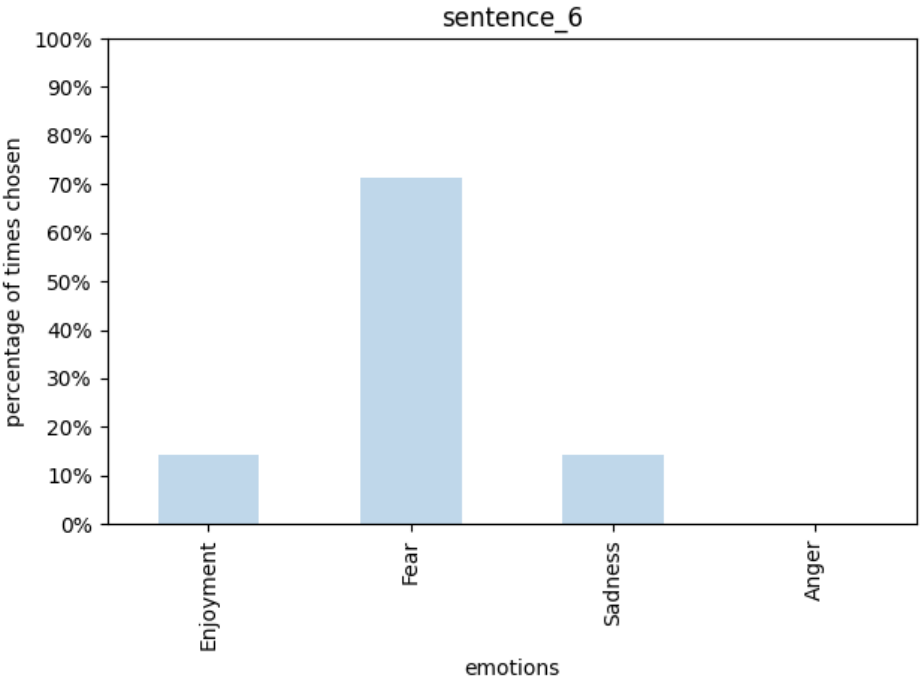

Figure B.50: Results for sentence_6 in user study 3.

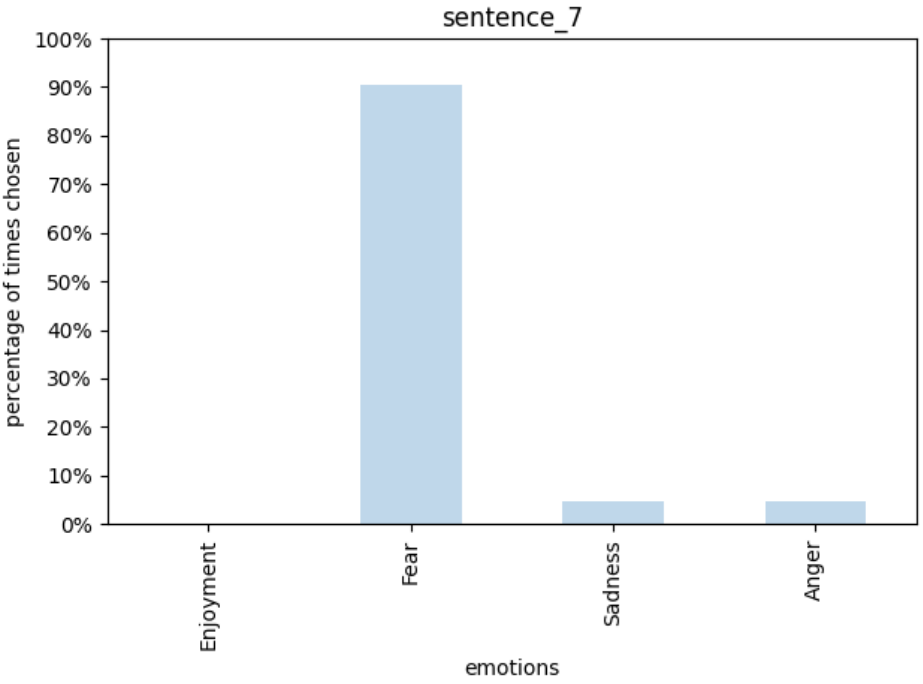

Figure B.51: Results for sentence_7 in user study 3.

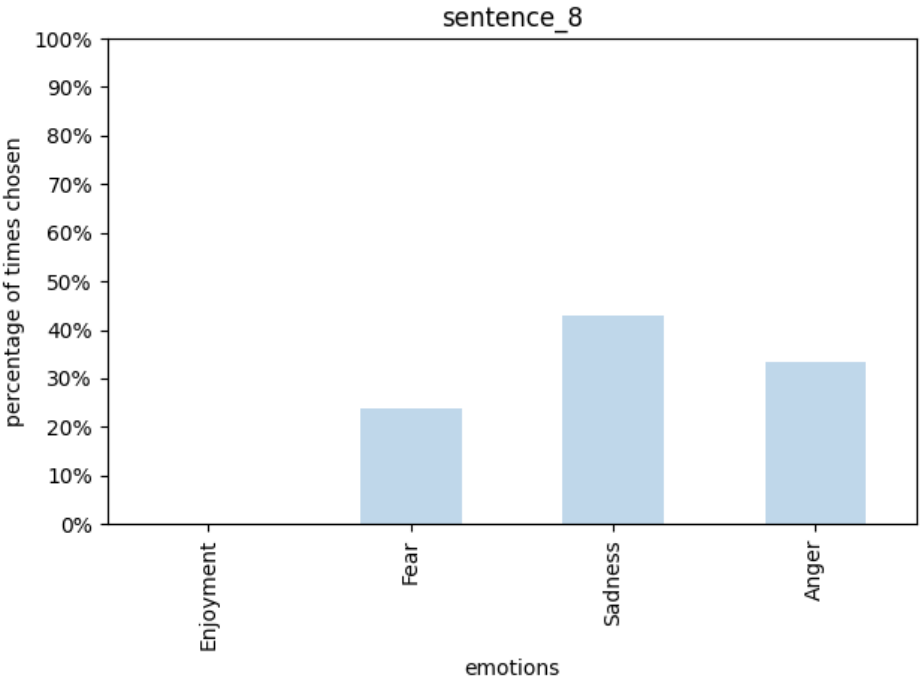

Figure B.52: Results for sentence_8 in user study 3.

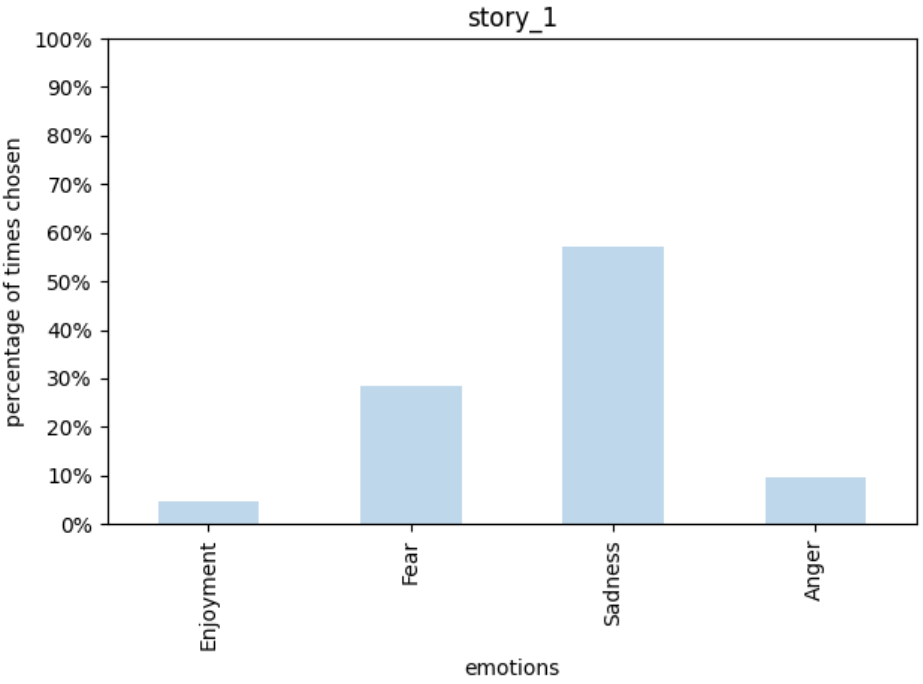

Figure B.53: Results for story_1 in user study 3.

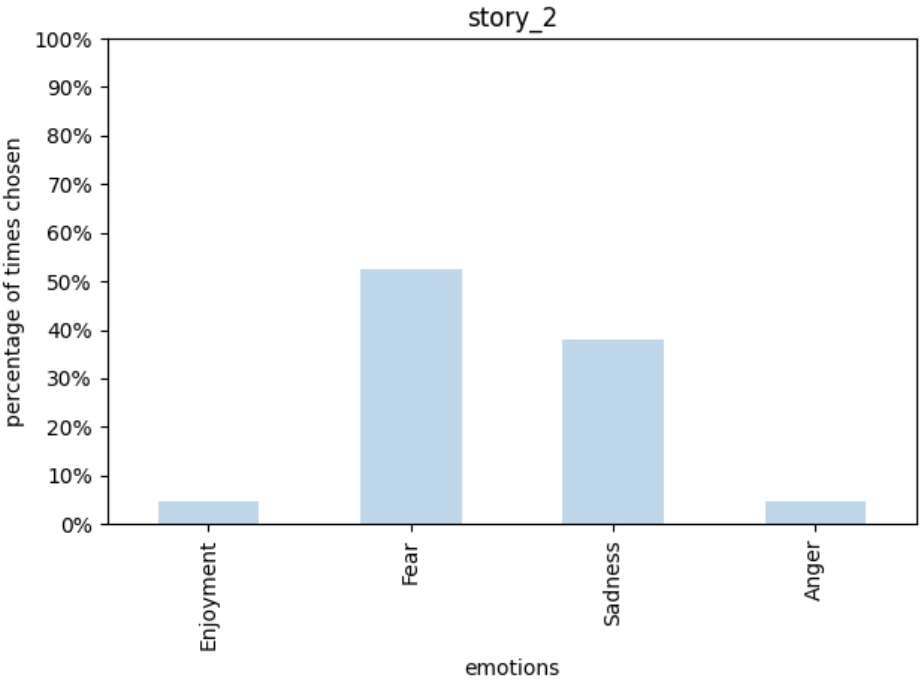

Figure B.54: Results for story_2 in user study 3.

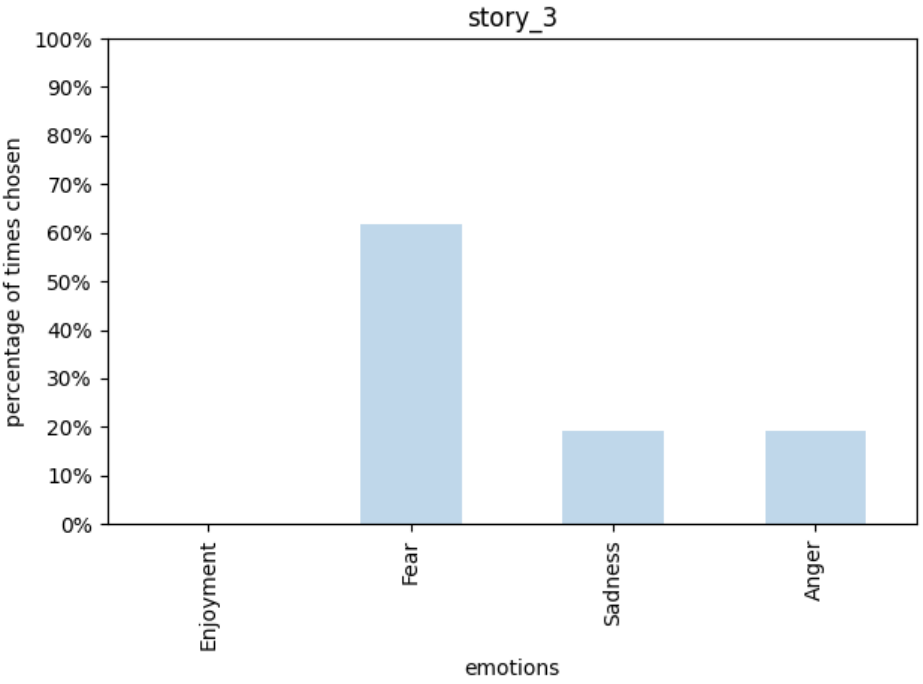

Figure B.55: Results for story_3 in user study 3.

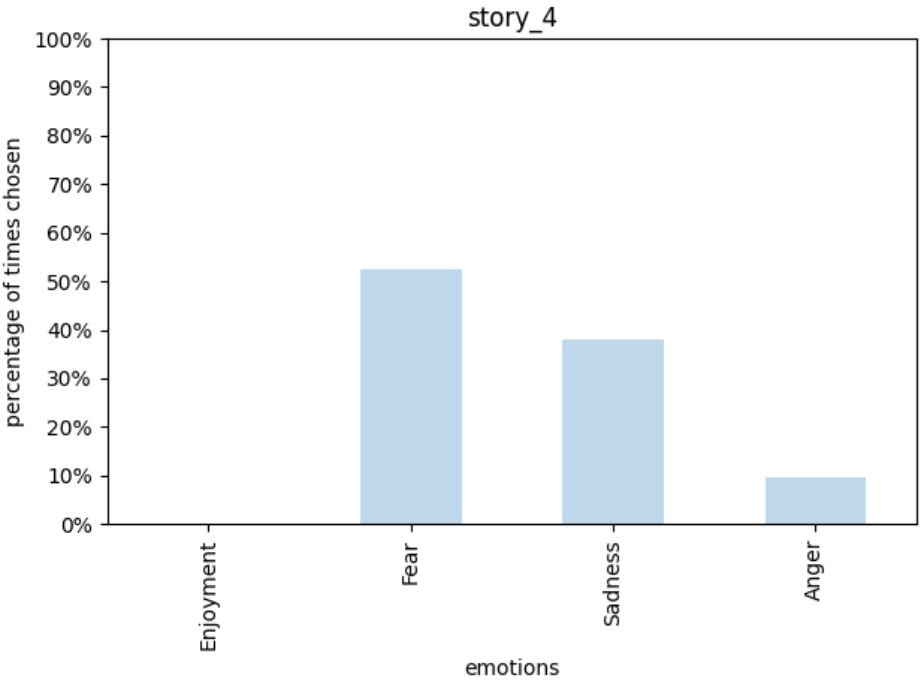

Figure B.56: Results for story_4 in user study 3.

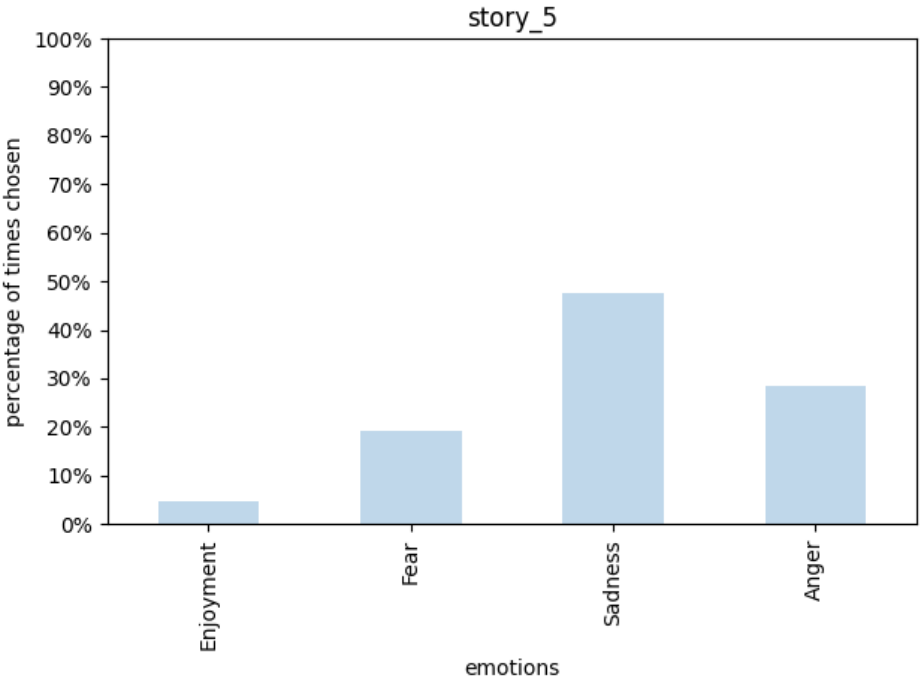

Figure B.57: Results for story_5 in user study 3.

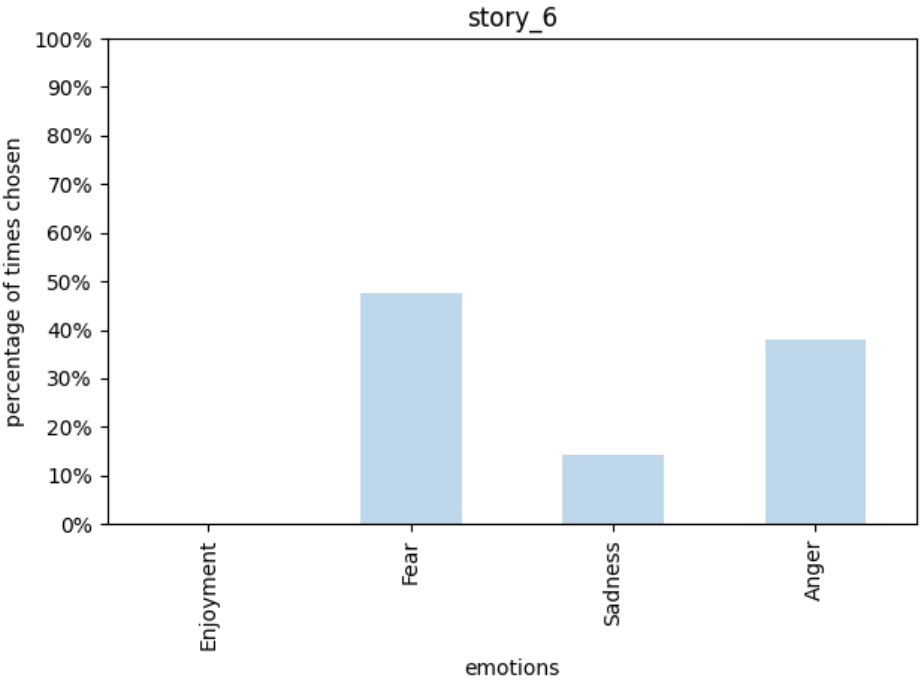

Figure B.58: Results for story_6 in user study 3.

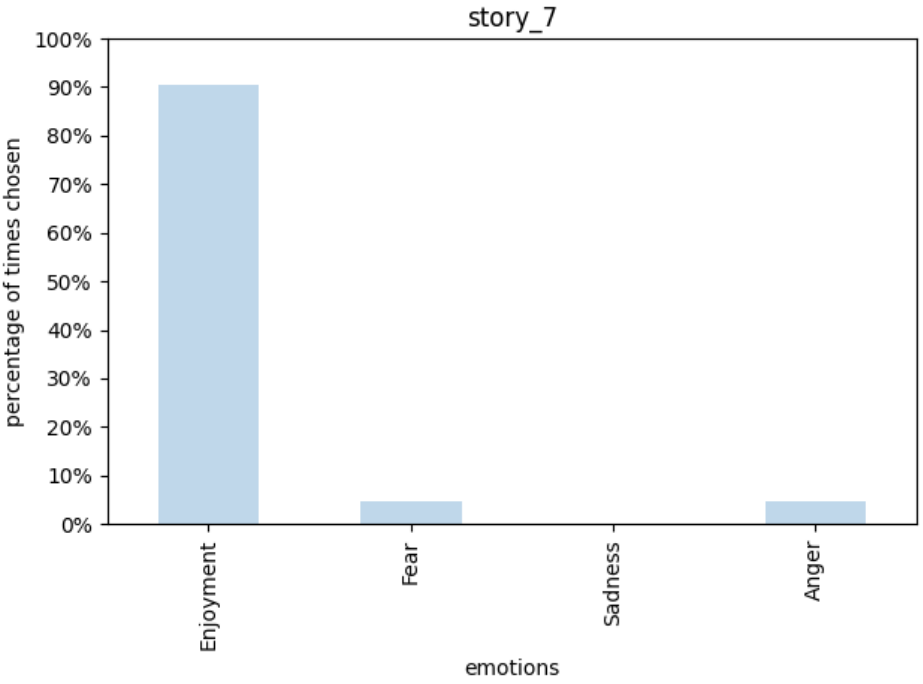

Figure B.59: Results for story_7 in user study 3.

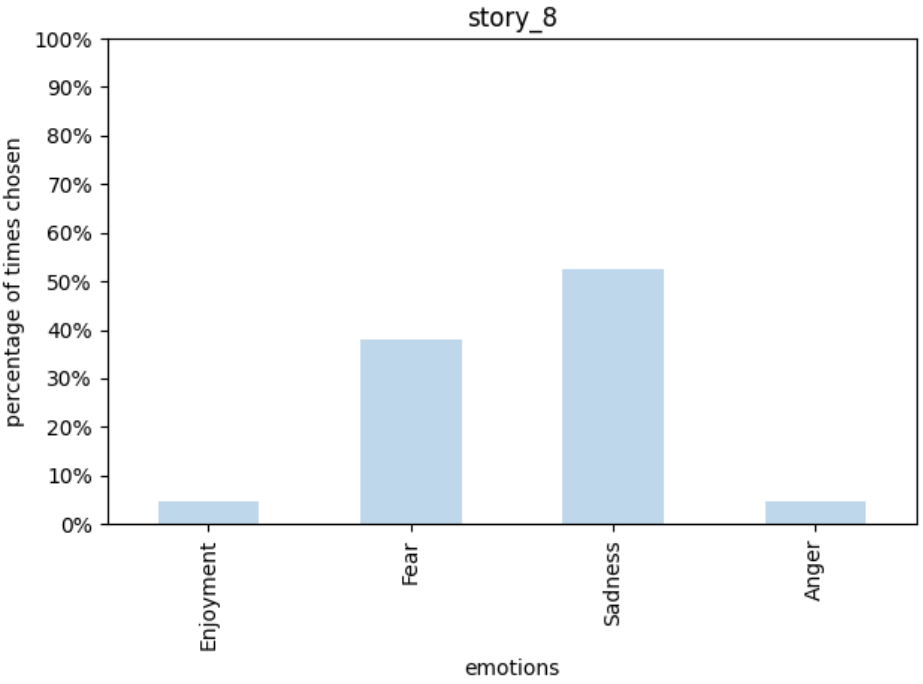

Figure B.60: Results for story_8 in user study 3.

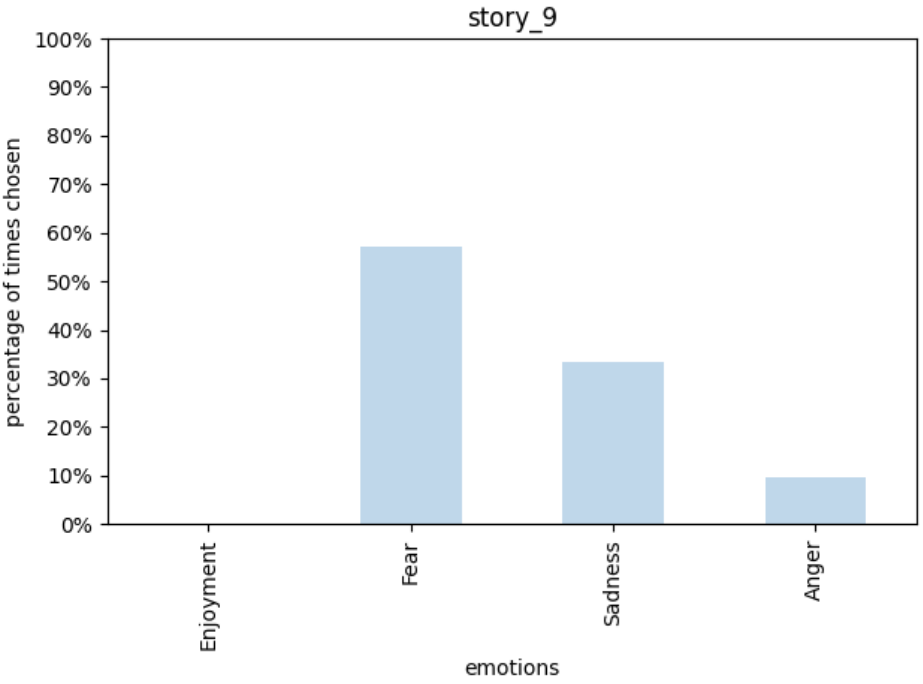

Figure B.61: Results for story_9 in user study 3.

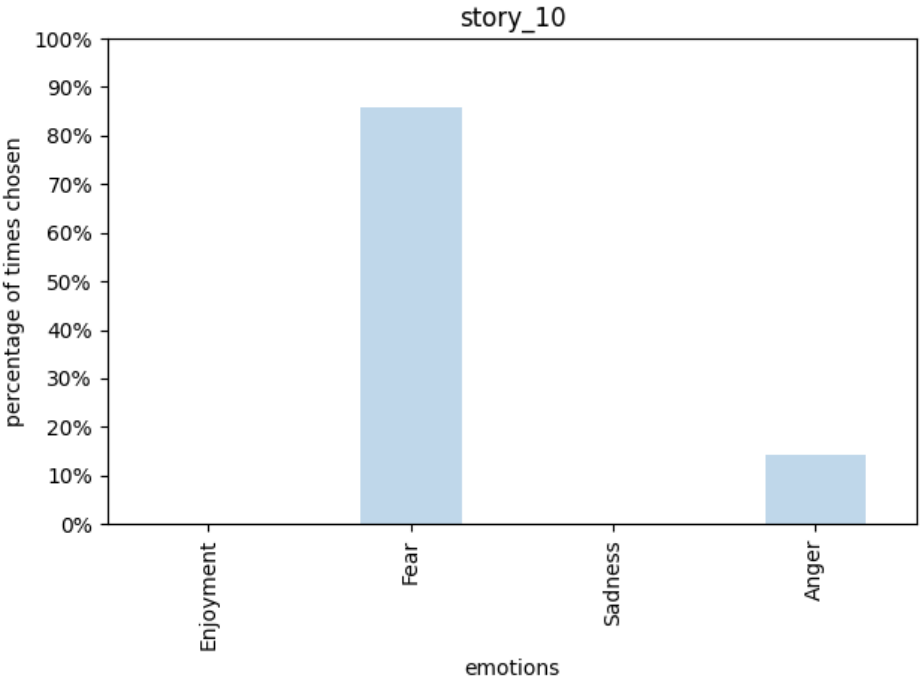

Figure B.62: Results for story_10 in user study 3.

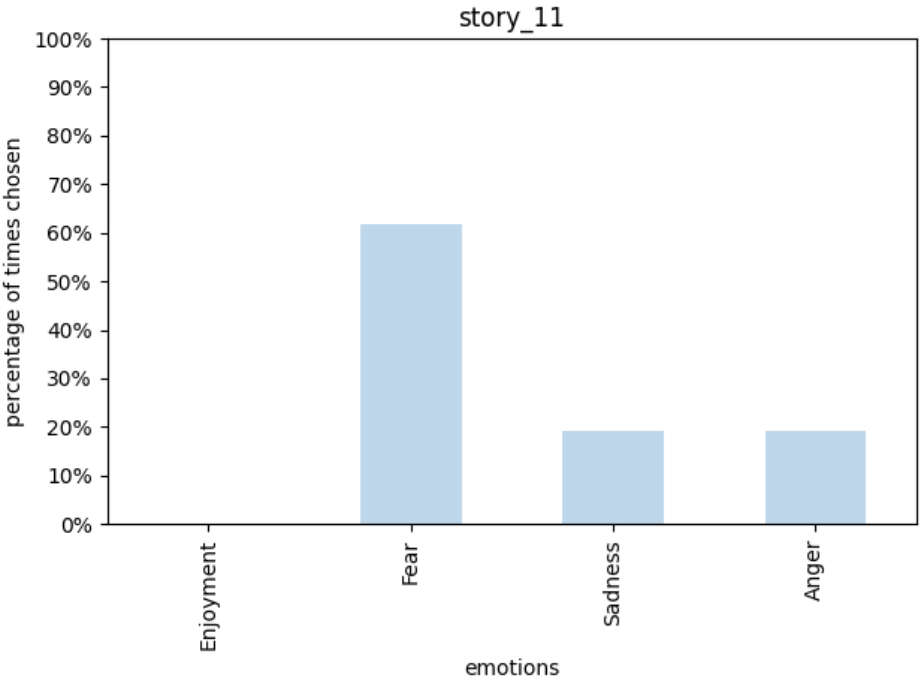

Figure B.63: Results for story_11 in user study 3.

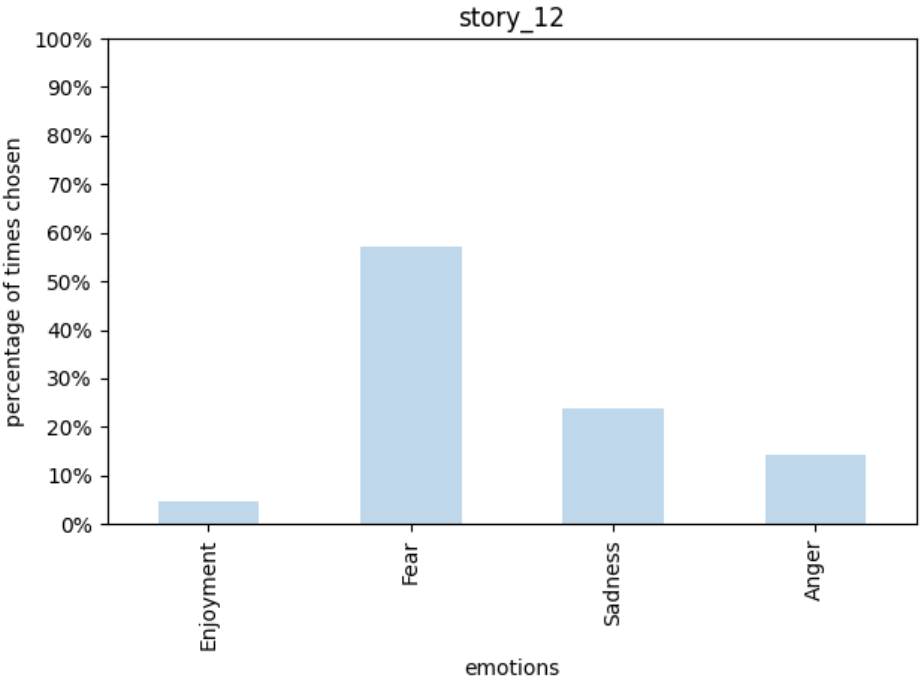

Figure B.64: Results for story_12 in user study 3.

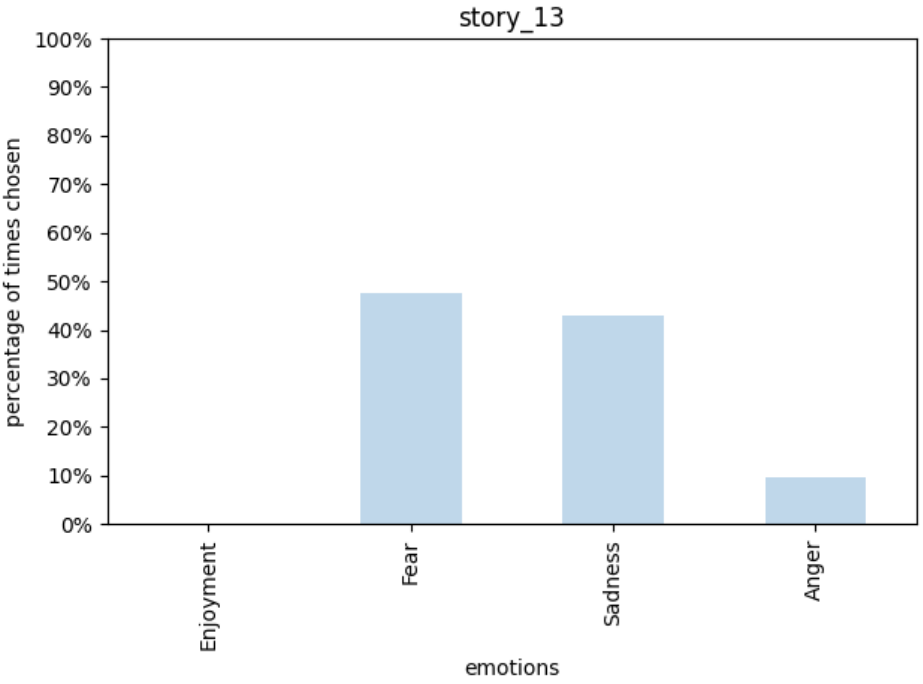

Figure B.65: Results for story_13 in user study 3.

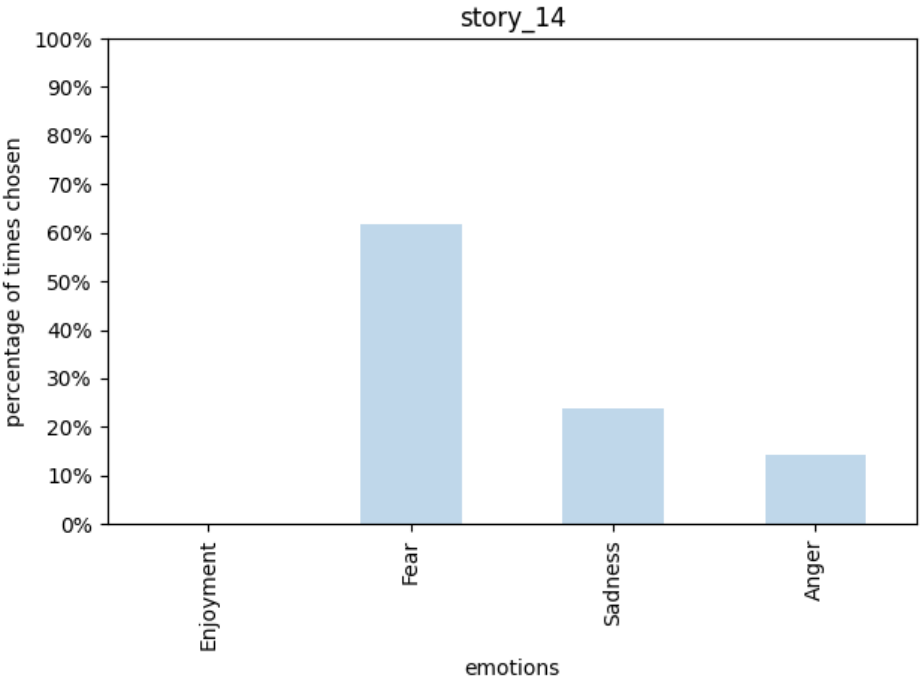

Figure B.66: Results for story_14 in user study 3.

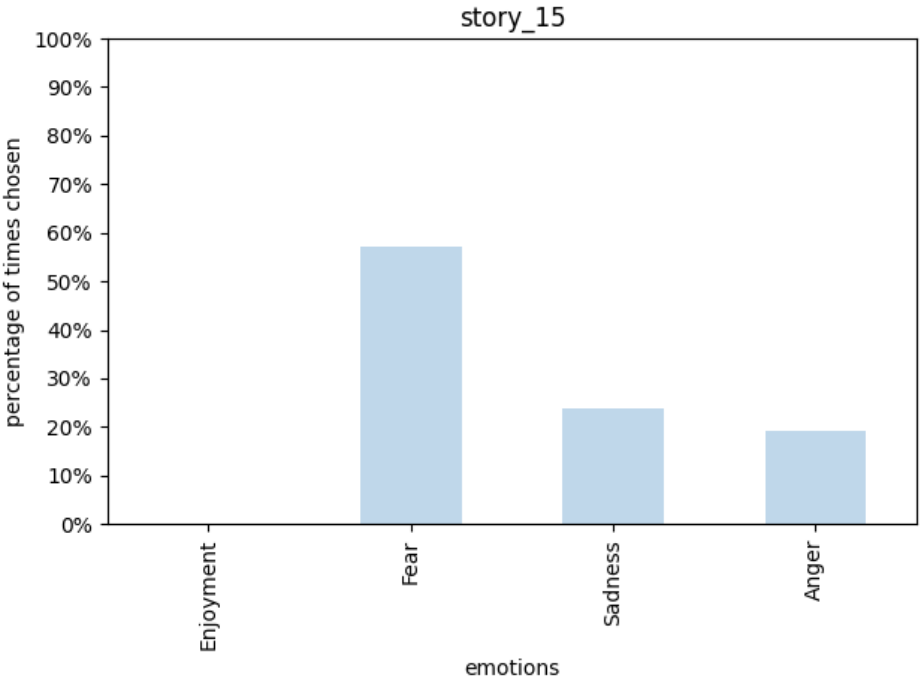

Figure B.67: Results for story_15 in user study 3.

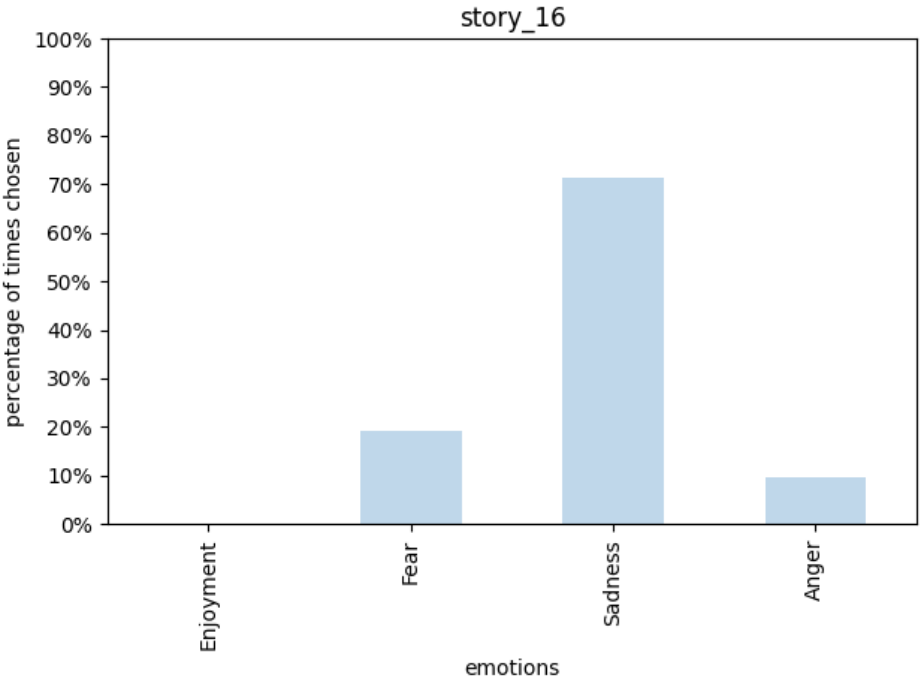

Figure B.68: Results for story_16 in user study 3.

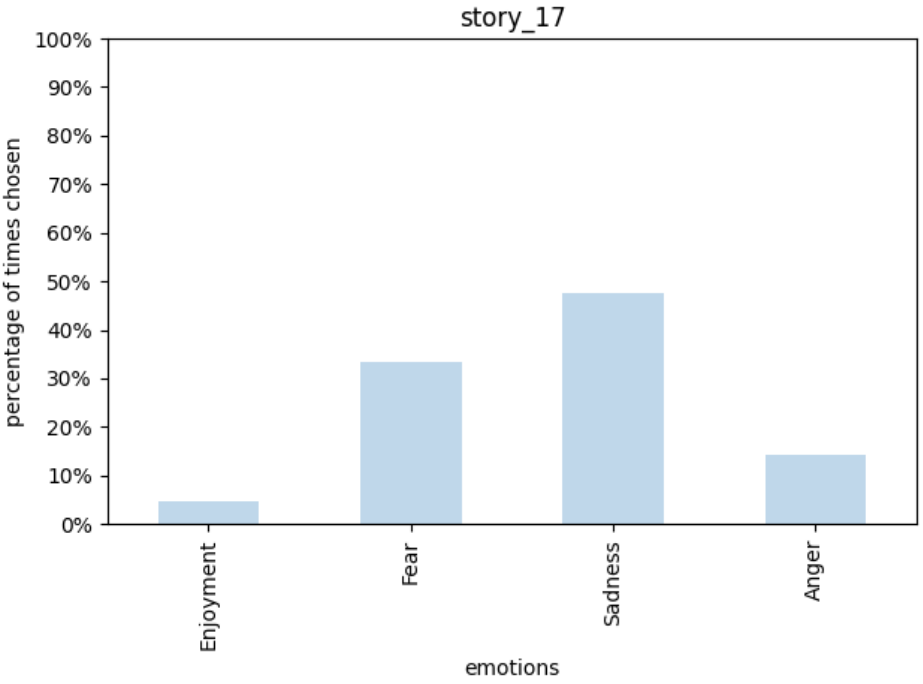

Figure B.69: Results for story_17 in user study 3.

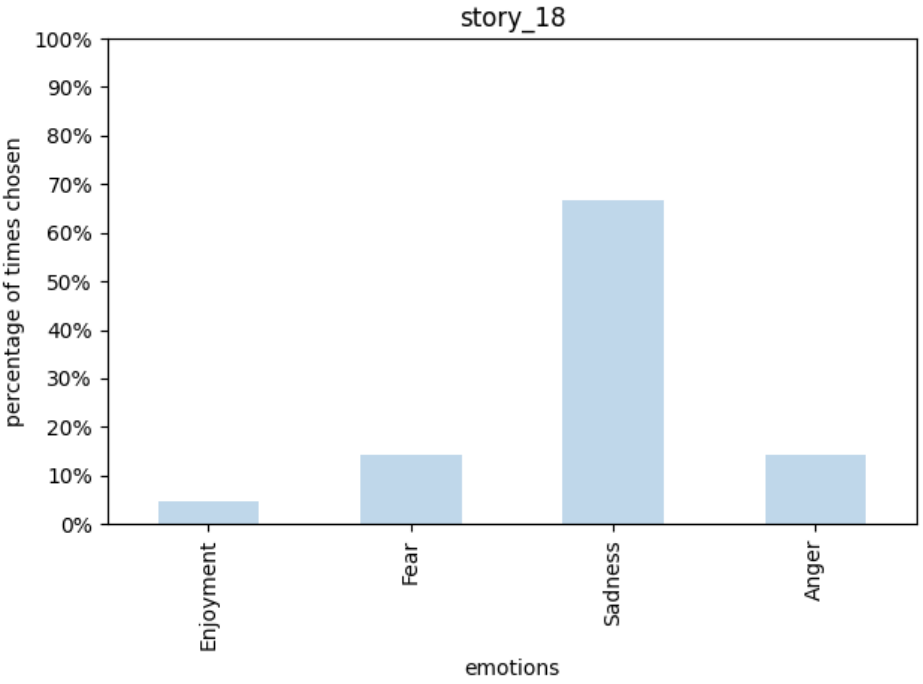

Figure B.70: Results for story_18 in user study 3.

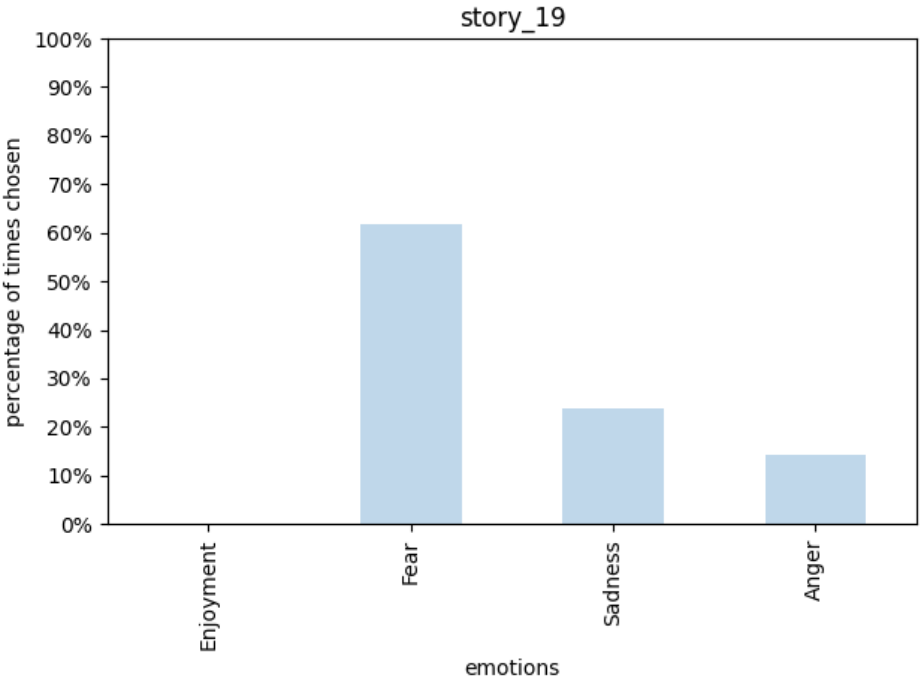

Figure B.71: Results for story_19 in user study 3.

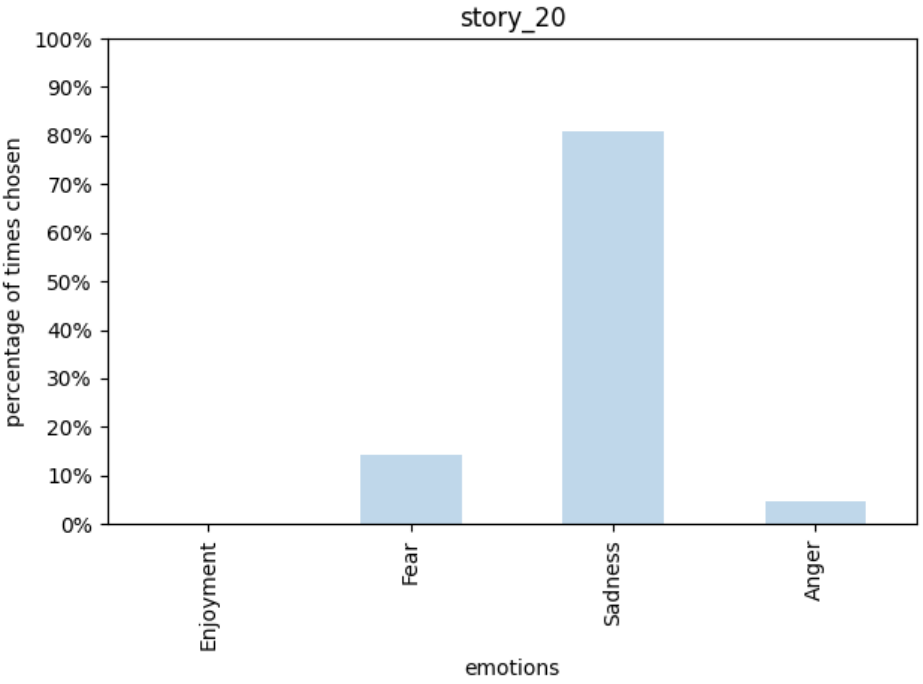

Figure B.72: Results for story_20 in user study 3.

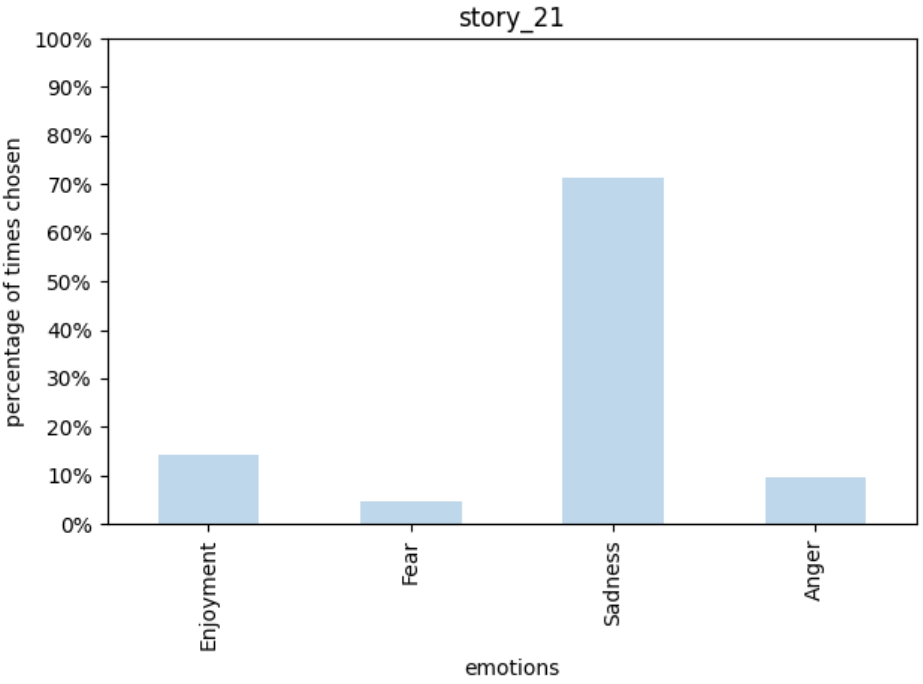

Figure B.73: Results for story_21 in user study 3.

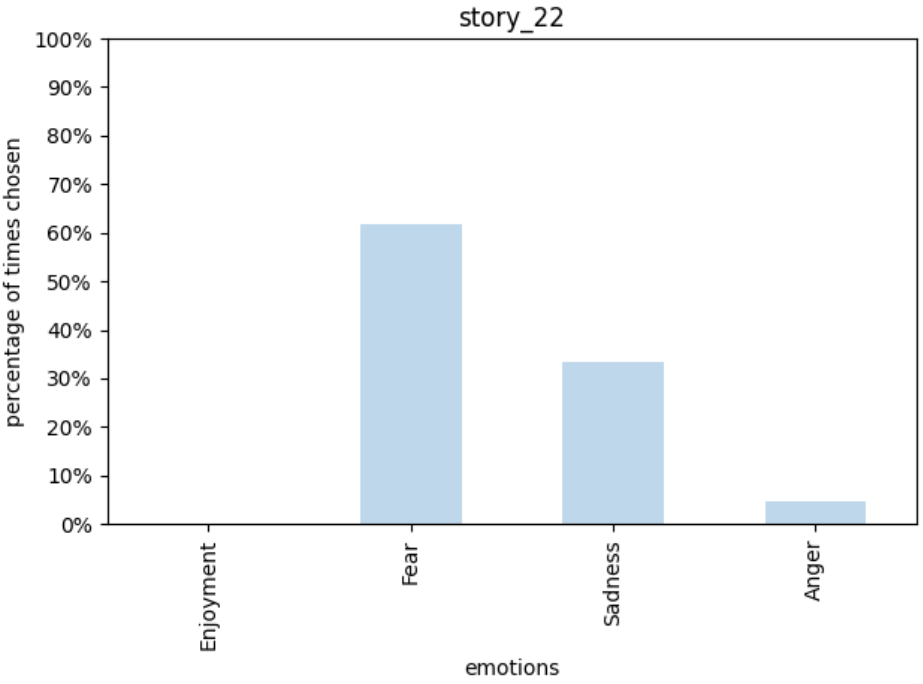

Figure B.74: Results for story_22 in user study 3.

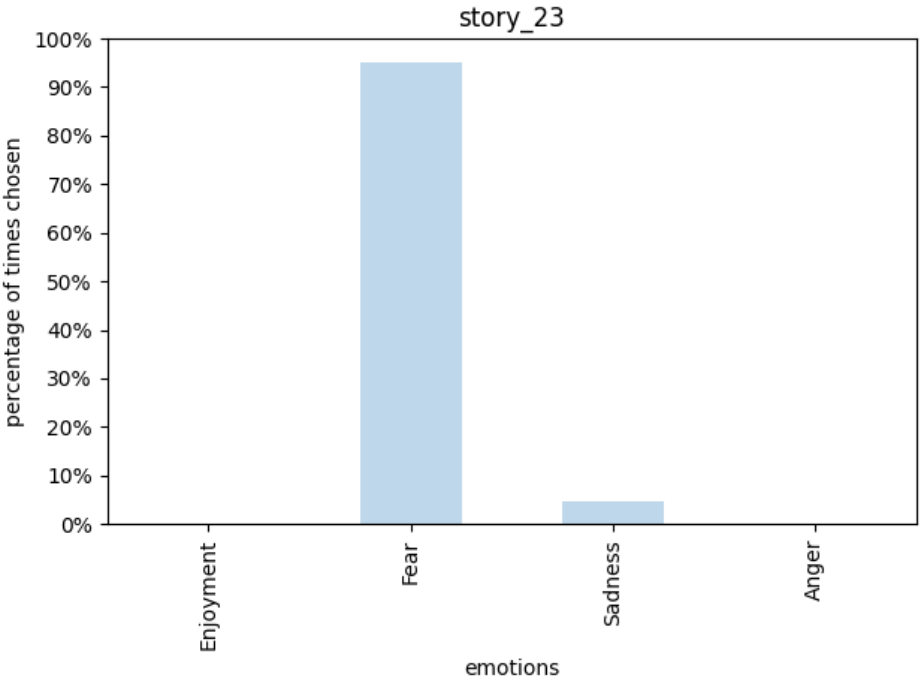

Figure B.75: Results for story_23 in user study 3.

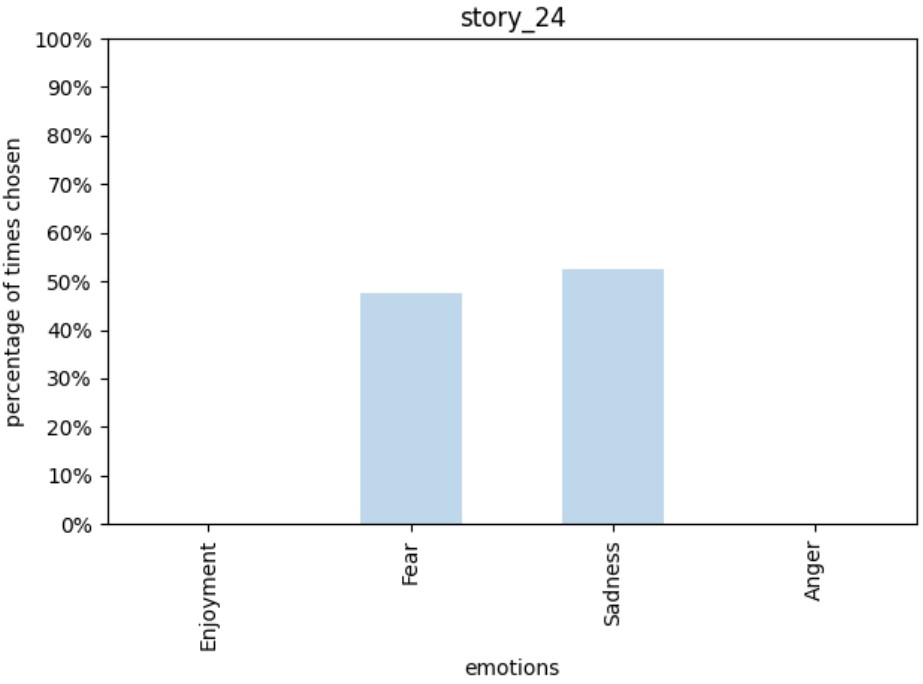

Figure B.76: Results for story_24 in user study 3.

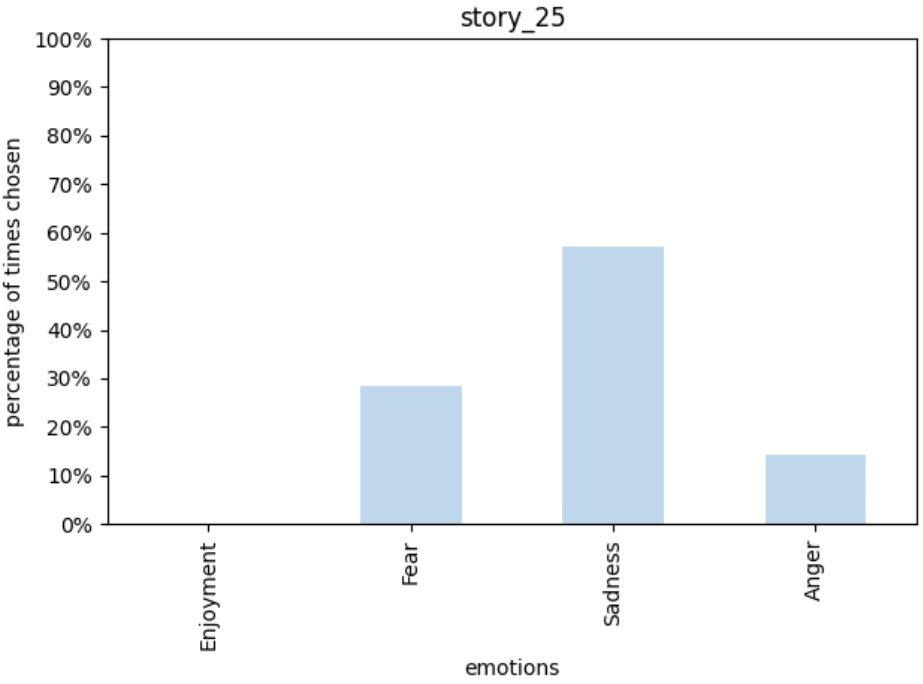

Figure B.77: Results for story_25 in user study 3.

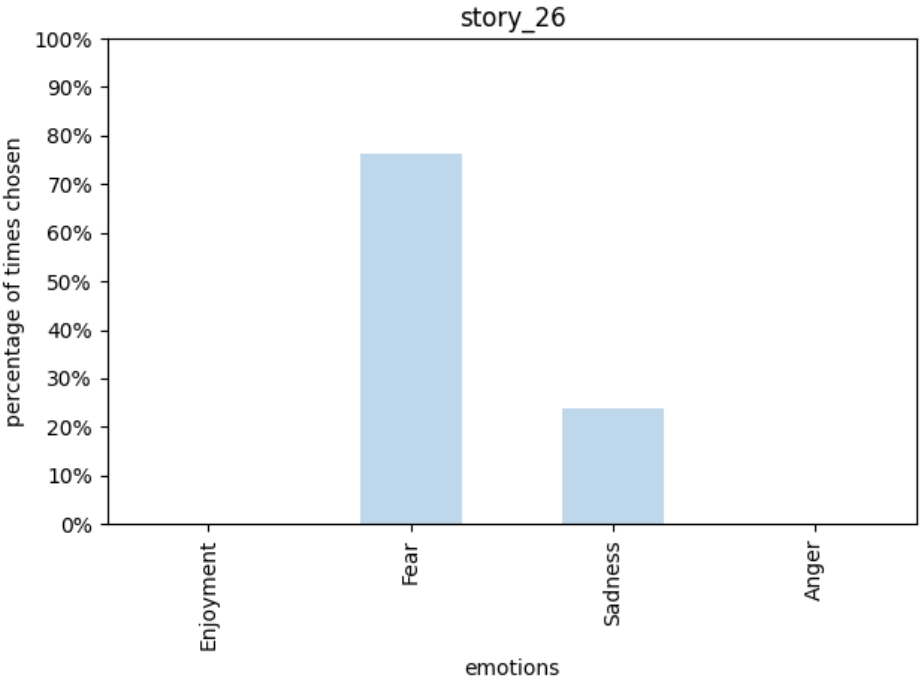

Figure B.78: Results for story_26 in user study 3.

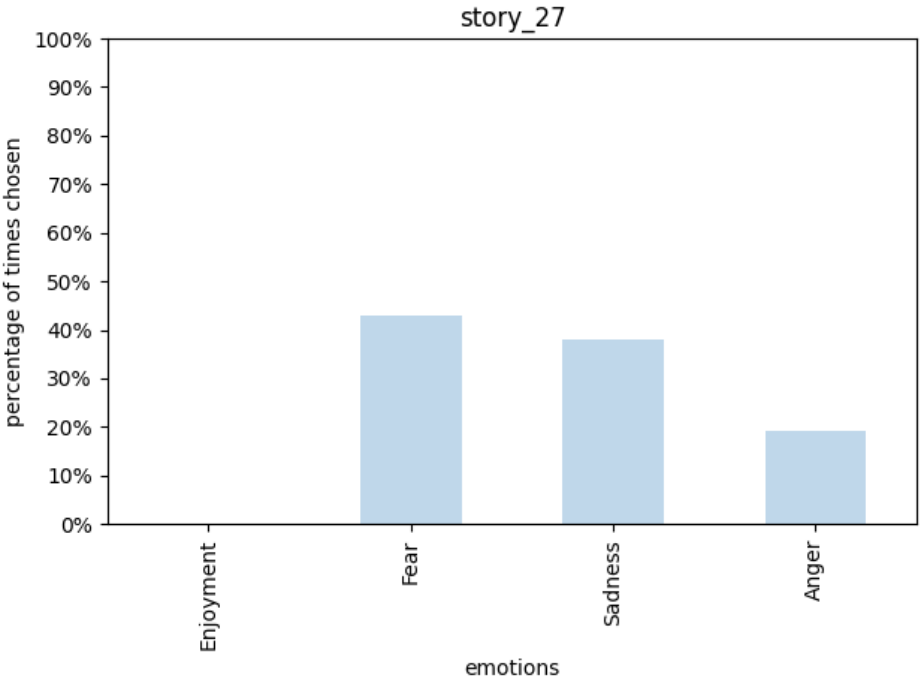

Figure B.79: Results for story_27 in user study 3.

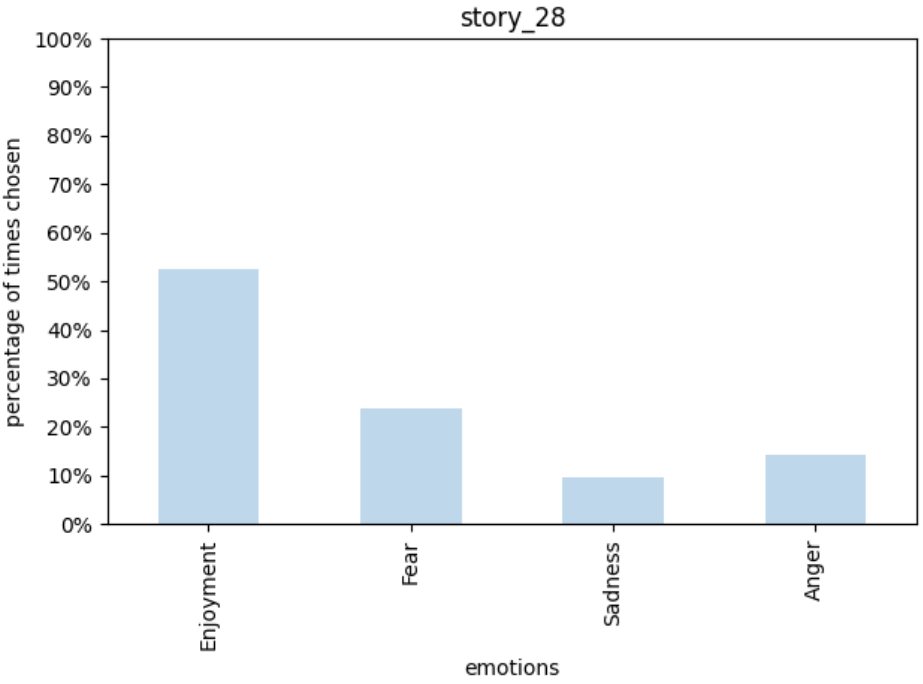

Figure B.80: Results for story_28 in user study 3.

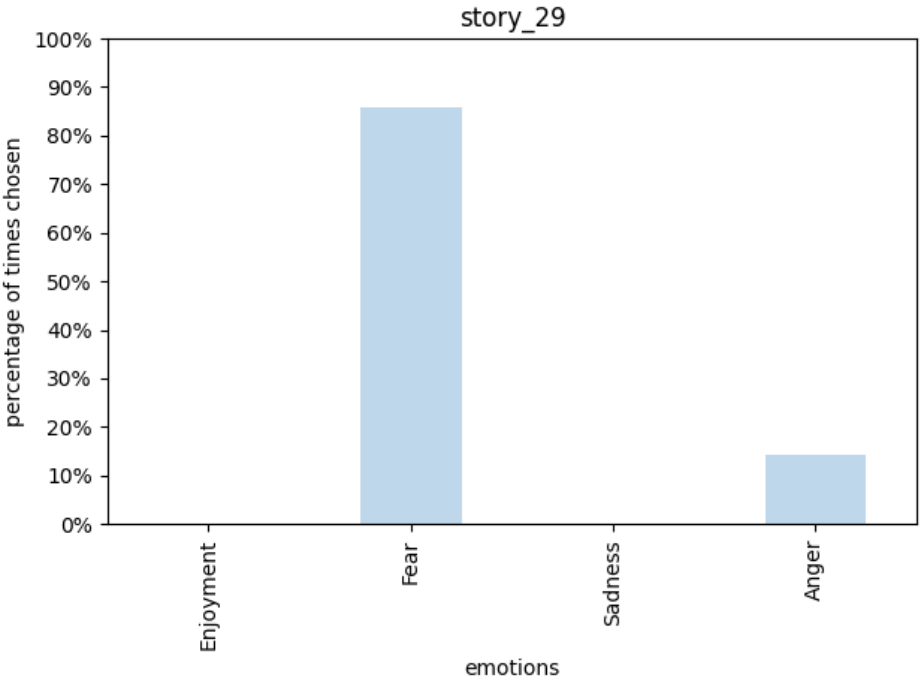

Figure B.81: Results for story_29 in user study 3.

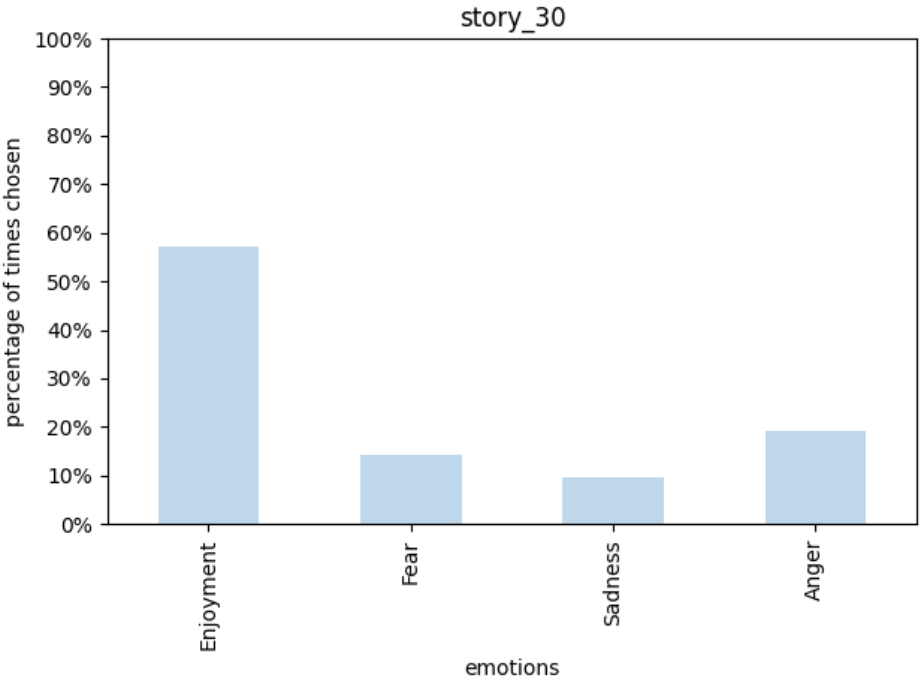

Figure B.82: Results for story_30 in user study 3.

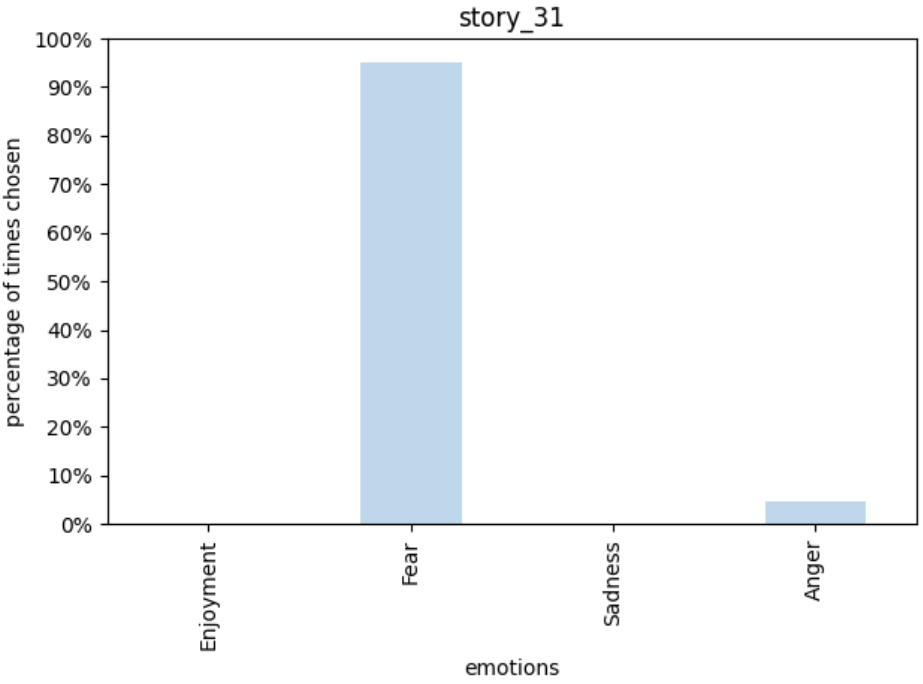

Figure B.83: Results for story_31 in user study 3.

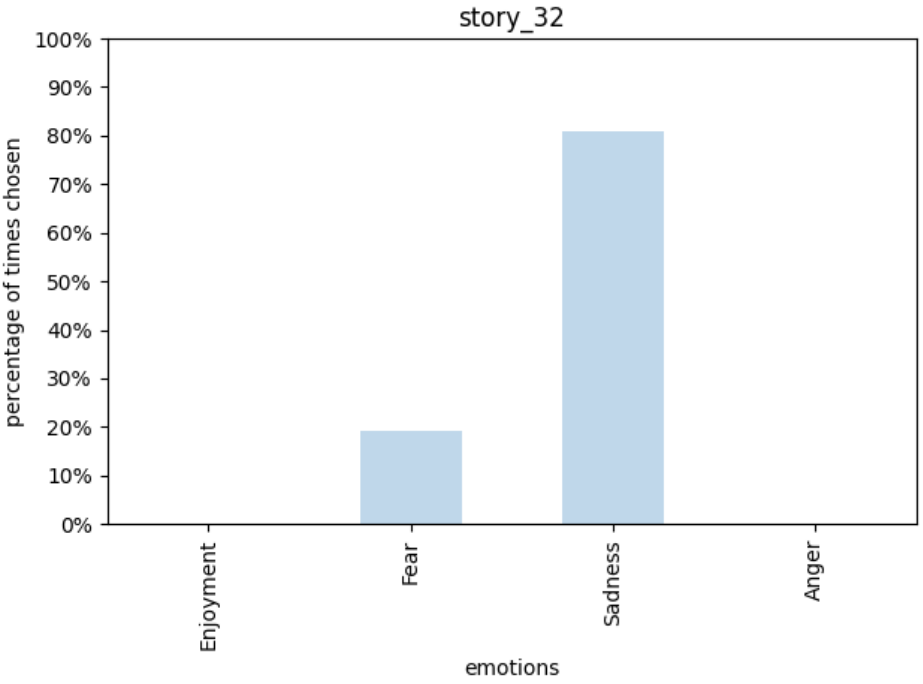

Figure B.84: Results for story_32 in user study 3.

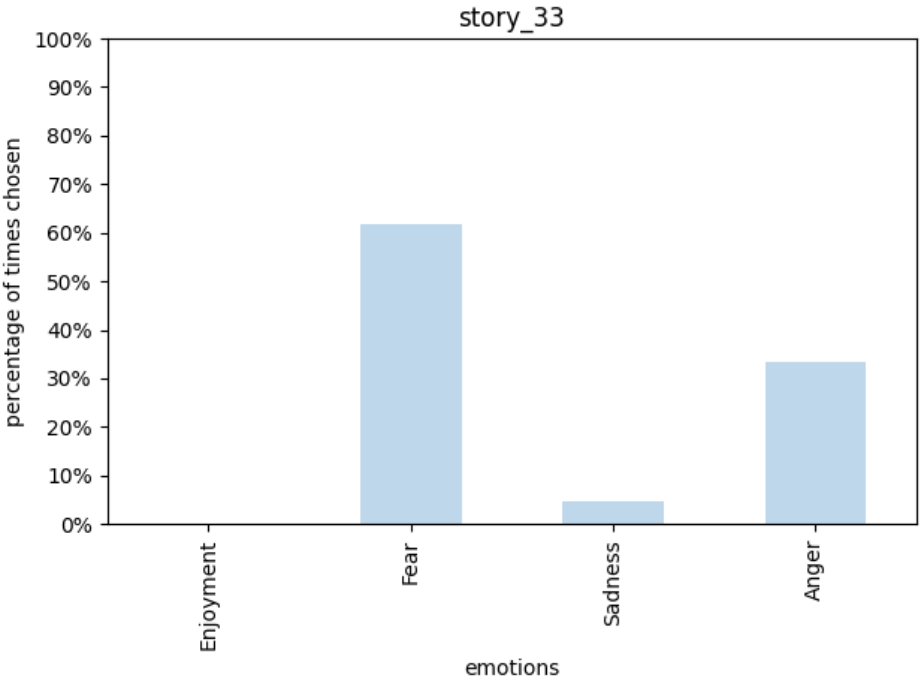

Figure B.85: Results for story_33 in user study 3.

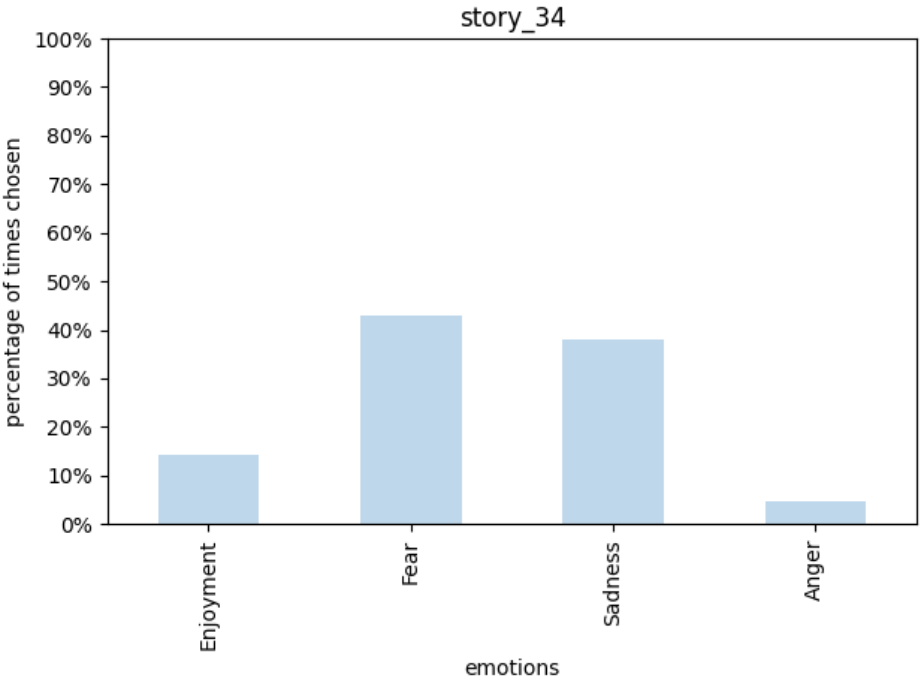

Figure B.86: Results for story_34 in user study 3.

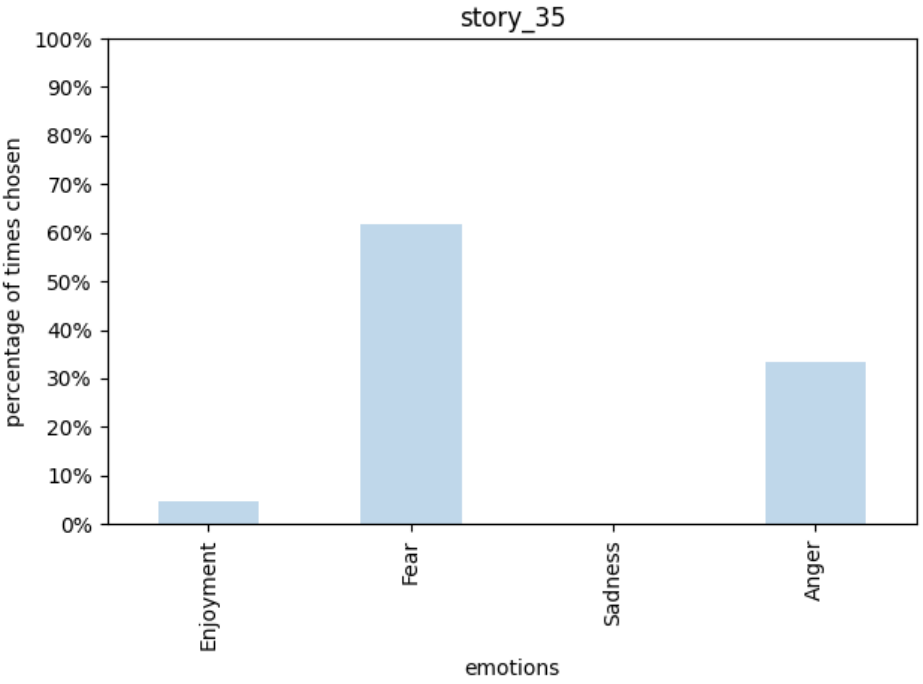

Figure B.87: Results for story_35 in user study 3.

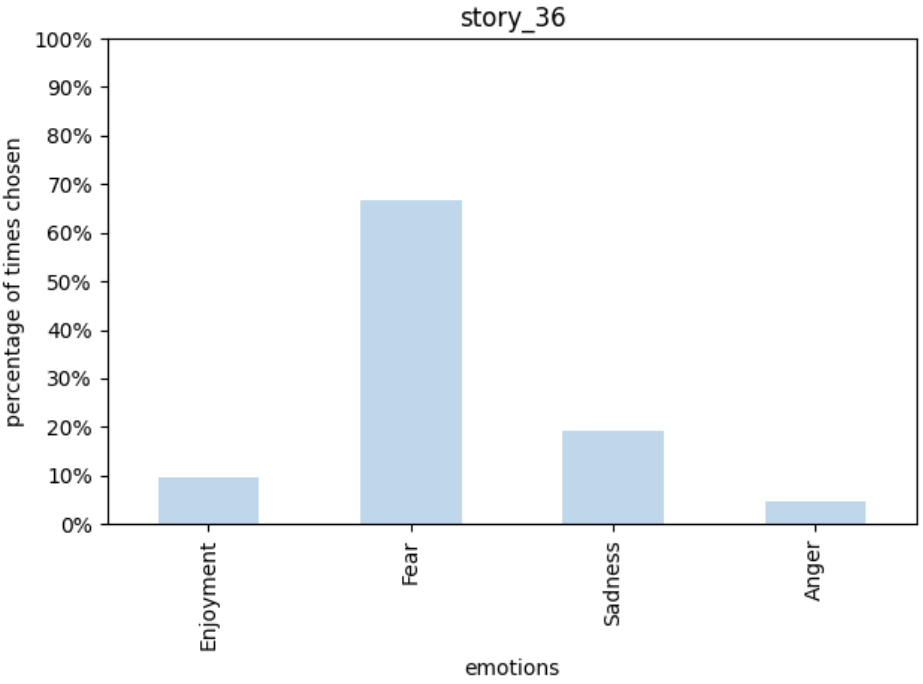

Figure B.88: Results for story_36 in user study 3.

Bibliography

[1] Martín Abadi et al. 'TensorFlow: A system for large-scale machine learning'. In: (2016). Publisher: arXiv Version Number: 2. DOI: 10.48550/ARXIV.1605.08695.

[2] Ahmad Abdellatif, Khaled Badran, Diego Elias Costa et al. 'A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering'. en. In: *arXiv:2012.02640 [cs]* (Dec. 2020). arXiv: 2012.02640.

[3] Ahmad Abdellatif, Khaled Badran and Emad Shihab. 'MSRBot: Using bots to answer questions from software repositories'. en. In: *Empirical Software Engineering* 25.3 (May 2020), pp. 1834–1863. ISSN: 1382-3256, 1573-7616. DOI: 10.1007/s10664-019-09788-5.

[4] Joyce A. Adams, Karen J. Farst and Nancy D. Kellogg. 'Interpretation of Medical Findings in Suspected Child Sexual Abuse: An Update for 2018'. en. In: *Journal of Pediatric and Adolescent Gynecology* 31.3 (June 2018), pp. 225–231. ISSN: 10833188. DOI: 10.1016/j.jpag.2017.12.011.

[5] Ane U. Albaek, Liv G. Kinn and Anne M. Milde. 'Walking Children Through a Minefield: How Professionals Experience Exploring Adverse Childhood Experiences'. en. In: *Qualitative Health Research* 28.2 (Jan. 2018), pp. 231–244. ISSN: 1049-7323, 1552-7557. DOI: 10.1177/1049732317734828.

[6] 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*. 2021, pp. 2–8.

[7] Tom Bocklisch et al. 'Rasa: Open Source Language Understanding and Dialogue Management'. In: *CoRR* abs/1712.05181 (2017). arXiv: 1712.05181.

[8] Bohn Stafleu van Loghum. 'Chatten met een robot helpt gepeste kinderen'. nl. In: *Jeugd en Co* 8.2 (Apr. 2014), pp. 5–5. ISSN: 1873-9164, 1876-6080. DOI: 10.1007/s12449-014-0024-5.

[9] Tom B. Brown et al. 'Language Models are Few-Shot Learners'. en. In: *arXiv:2005.14165 [cs]* (July 2020). arXiv: 2005.14165.

[10] Ray Bull and Becky Milne. 'Attempts to Improve the Police Interviewing of Suspects'. In: *Interrogations, Confessions, and Entrapment*. Ed. by G. Daniel Lassiter. Boston, MA: Springer US, 2004, pp. 181–196. ISBN: 978-0-387-38598-3. DOI: 10.1007/978-0-387-38598-3_8.

[11] Tanja Bunk et al. 'DIET: Lightweight Language Understanding for Dialogue Systems'. In: (2020). Publisher: arXiv Version Number: 3. DOI: 10.48550/ARXIV.2004.09936.

[12] Massimo Canonico and Luigi De Russis. 'A Comparison and Critique of Natural Language Understanding Tools'. en. In: *CLOUD COMPUTING* (2018), p. 7.

[13] Hannah Cassidy, Lucy Akehurst and Julie Cherryman. 'Police Interviewers' Perceptions of Child Credibility in Forensic Investigations'. en. In: *Psychiatry, Psychology and Law* 27.1 (Jan. 2020), pp. 61–80. ISSN: 1321-8719, 1934-1687. DOI: 10.1080/13218719.2019.1687044.

[14] Dr Ann Cavoukian. 'Information & Privacy Commissioner of Ontario'. en. In: (Jan. 2009), p. 6.

[15] Ann-Christin Cederborg et al. 'Investigative interviews of child witnesses in Sweden'. en. In: *Child Abuse & Neglect* 24.10 (Oct. 2000), pp. 1355–1361. ISSN: 01452134. DOI: 10.1016/S0145-2134(00)00183-6.

[16] 'Chatbots in the fight against the COVID-19 pandemic'. English. In: *npj Digital Medicine* 3 (May 2020), pp. 1–4. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0280-0.

[17] *Child maltreatment*. June 2020.

[18] Hayley M. D. Cleary and Todd C. Warner. 'Police training in interviewing and interrogation methods: A comparison of techniques used with adult and juvenile suspects.' en. In: *Law and Human Behavior* 40.3 (2016), pp. 270–284. ISSN: 1573-661X, 0147-7307. DOI: 10.1037/lhb0000175.

[19] D. E. Comer et al. 'Computing as a discipline'. en. In: *Communications of the ACM* 32.1 (Jan. 1989). Ed. by Peter J. Denning, pp. 9–23. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/63238.63239.

[20] Thomas H. Cormen, ed. *Introduction to algorithms*. 3rd ed. OCLC: ocn311310321. Cambridge, Mass: MIT Press, 2009. ISBN: 978-0-262-03384-8.

[21] Janet Currie and Cathy Spatz Widom. 'Long-Term Consequences of Child Abuse and Neglect on Adult Economic Well-Being'. en. In: *Child Maltreatment* 15.2 (May 2010), pp. 111–120. ISSN: 1077-5595, 1552-6119. DOI: 10.1177/1077559509355316.

[22] Avisha Das and Rakesh M. Verma. 'Can Machines Tell Stories? A Comparative Study of Deep Neural Language Models and Metrics'. en. In: *IEEE Access* 8 (2020), pp. 181258–181292. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3023421.

[23] Florence Davey-Attlee and Isa Soares. *The fake news machine: Inside a town gearing up for 2020*. English. CNN. Sept. 2017.

[24] Jacob Devlin et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805.

[25] Tone Dyrhaug and Unni Beate Grebstad. *Child welfare*. July 2020.

[26] Marleen Easton, ed. *Multiple community policing: Hoezo?* dut. Reeks "Samenleving en toekomst". Gent: Academia Press [u.a.], 2009. ISBN: 978-90-382-1493-1.

[27] Paul Ekman and Wallace V. Friesen. 'A new pan-cultural facial expression of emotion'. en. In: *Motivation and Emotion* 10.2 (June 1986), pp. 159–168. ISSN: 0146-7239, 1573-6644. DOI: 10.1007/BF00992253.

[28] Paul Ekman and Karl G. Heider. 'The universality of a contempt expression: A replication'. en. In: *Motivation and Emotion* 12.3 (Sept. 1988), pp. 303–308. ISSN: 0146-7239, 1573-6644. DOI: 10.1007/BF00993116.

[29] Robert Faris et al. 'Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election'. en. In: (2016), p. 142.

[30] Christiane Fellbaum, ed. *WordNet: An Electronic Lexical Database*. en. The MIT Press, 1998. ISBN: 978-0-262-27255-1. DOI: 10.7551/mitpress/7287.001.0001.

[31] Gail S. Goodman and Annika Melinder. 'Child witness research and forensic interviews of young children: A review'. en. In: *Legal and Criminological Psychology* 12.1 (Feb. 2007), pp. 1–19. ISSN: 13553259. DOI: 10.1348/135532506X156620.

[32] Google. *DialogFlow - Documentation - Editions*. Documentation guides.

[33] Sara E. Gorman and Jack M. Gorman. *Denying to the Grave: Why We Ignore the Facts That Will Save Us*. Aug. 2016. ISBN: 978-0-19-939660-3.

[34] Eric Gregori. 'Evaluation of Modern Tools for an OMSCS Advisor Chatbot'. en. In: (Aug. 2017), p. 50.

[35] Simeng Gu et al. 'A Model for Basic Emotions Using Observations of Behavior in Drosophila'. en. In: *Frontiers in Psychology* 10 (Apr. 2019), p. 781. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00781.

[36] Wael H.Gomaa and Aly A. Fahmy. 'A Survey of Text Similarity Approaches'. en. In: *International Journal of Computer Applications* 68.13 (Apr. 2013), pp. 13–18. ISSN: 09758887. DOI: 10.5120/11638-7118.

[37] Joshua K. Hartshorne and Laura T. Germine. 'When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span'. en. In: *Psychological Science* 26.4 (Apr. 2015), pp. 433–443. ISSN: 0956-7976, 1467-9280. DOI: 10.1177/0956797614567339.

[38] Syed Zohaib Hassan et al. 'Towards an AI-Driven Talking Avatar in Virtual Reality for Investigative Interviews of Children'. In: *Proceedings of the Workshop on Game Systems (GameSys' 22)*. 2022.

[39] Pengcheng He et al. 'DeBERTa: Decoding-enhanced BERT with Disentangled Attention'. en. In: *arXiv:2006.03654 [cs]* (Oct. 2021). arXiv: 2006.03654.

[40] Julia Hirschberg and Christopher D Manning. 'Advances in natural language processing'. en. In: *ARTIFICIAL INTELLIGENCE* 349.6245 (July 2015), p. 7.

[41] Arlie Russel Hochschild. *The managed heart: commercialization of human feeling*. eng. Updated ed. Berkeley, Calif. London: University of California Press, 2012. ISBN: 978-0-520-27294-1.

[42] Ari Holtzman et al. 'The Curious Case of Neural Text Degeneration'. en. In: *arXiv:1904.09751 [cs]* (Feb. 2020). arXiv: 1904.09751.

[43] Scott Huffman. *Making Conversational Interfaces Easier to Build*. English. Sept. 2016.

[44] Barrie Irving. *Police Interrogation: The Psychological Approach*. en. 1980. ISBN: 0-11-730122-1.

[45] Miriam Johnson et al. 'Best Practice Recommendations Still Fail to Result in Action: A National 10-Year Follow-up Study of Investigative Interviews in CSA Cases: Follow-up study of investigative interviews'. en. In: *Applied Cognitive Psychology* 29.5 (Sept. 2015), pp. 661–668. ISSN: 08884080. DOI: 10.1002/acp.3147.

[46] Dana L. Joseph and Daniel A. Newman. 'Emotional intelligence: An integrative meta-analysis and cascading model.' en. In: *Journal of Applied Psychology* 95.1 (Jan. 2010), pp. 54–78. ISSN: 1939-1854, 0021-9010. DOI: 10.1037/a0017286.

[47] Chien-Hao Kao, Chih-Chieh Chen and Yu-Tza Tsai. 'Model of Multi-turn Dialogue in Emotional Chatbot'. en. In: *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. Kaohsiung, Taiwan: IEEE, Nov. 2019, pp. 1–5. ISBN: 978-1-72814-666-9. DOI: 10.1109/TAAI48200.2019.8959855.

[48] Eda Kavlakoglu. *NLP vs. NLU vs. NLG: the differences between three natural language processing concepts*. English. IBM. Nov. 2020.

[49] Patricia K. Kerig et al. 'Numbing of Positive, Negative, and General Emotions: Associations With Trauma Exposure, Posttraumatic Stress, and Depressive Symptoms Among Justice-Involved Youth: Numbing of Positive, Negative, or General Emotions'. en. In: *Journal of Traumatic Stress* 29.2 (Apr. 2016), pp. 111–119. ISSN: 08949867. DOI: 10.1002/jts.22087.

[50] Nitish Shirish Keskar et al. 'CTRL: A Conditional Transformer Language Model for Controllable Generation'. en. In: *arXiv:1909.05858 [cs]* (Sept. 2019). arXiv: 1909.05858.

[51] Akbir Khan. 'Latent Racial Bias – Evaluating Racism in Police Stop-and-Searches'. en. In: *arXiv:2005.13463 [stat]* (May 2020). arXiv: 2005.13463.

[52] Terry K. Koo and Mae Y. Li. 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research'. en. In: *Journal of Chiropractic Medicine* 15.2 (June 2016), pp. 155–163. ISSN: 15563707. DOI: 10.1016/j.jcm.2016.02.012.

[53] Chethan Kumar. *NLP vs NLU vs NLG (Know what you are trying to achieve) NLP engine (Part-1)*. English. Towards Data Science. Sept. 2018.

[54] Alfirna Rizqi Lahitani, Adhistya Erna Permanasari and Noor Akhmad Setiawan. 'Cosine similarity to determine similarity measure: Study case in online essay assessment'. en. In: *2016 4th International Conference on Cyber and IT Service Management*. Bandung, Indonesia: IEEE, Apr. 2016, pp. 1–6. ISBN: 978-1-4673-8443-8. DOI: 10.1109/CITSM.2016.7577578.

[55] Michael E Lamb et al. *Children's testimony: A handbook of psychological research and forensic practice*. Vol. 53. John Wiley & Sons, 2011.

[56] Michael E. Lamb. 'Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink?' en. In: *American Psychologist* 71.8 (Nov. 2016), pp. 710–718. ISSN: 1935-990X, 0003-066X. DOI: 10.1037/amp0000039.

[57] Sungbok Lee, Alexandros Potamianos and Shrikanth Narayanan. 'Acoustics of children's speech: Developmental changes of temporal and spectral parametersa)'. en. In: (1999), p. 14.

[58] Jochen L. Leidner and Vassilis Plachouras. 'Ethical by Design: Ethics Best Practices for Natural Language Processing'. en. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 30–40. DOI: 10.18653/v1/W17-1604.

[59] Yoav Levine et al. 'SenseBERT: Driving Some Sense into BERT'. en. In: *arXiv:1908.05646 [cs]* (May 2020). arXiv: 1908.05646.

[60] Mike Lewis et al. 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension'. en. In: *arXiv:1910.13461 [cs, stat]* (Oct. 2019). arXiv: 1910.13461.

[61] Jason Liao et al. *Wit.ai GitHub*. May 2021.

[62] Yinhan Liu et al. 'RoBERTa: A Robustly Optimized BERT Pre-training Approach'. en. In: *arXiv:1907.11692 [cs]* (July 2019). arXiv: 1907.11692.

[63] Mélanie Loiseau. 'Physical abuse of young children during the COVID-19 pandemic: Alarming increase in the relative frequency of hospitalizations during the lockdown period'. en. In: *Child Abuse* (2021), p. 8.

[64] Gale M. Lucas et al. 'It's only a computer: Virtual humans increase willingness to disclose'. en. In: *Computers in Human Behavior* 37 (Aug. 2014), pp. 94–100. ISSN: 07475632. DOI: 10.1016/j.chb.2014.04.043.

[65] David Matsumoto. 'More evidence for the universality of a contempt expression'. en. In: *Motivation and Emotion* 16.4 (Dec. 1992), pp. 363–368. ISSN: 0146-7239, 1573-6644. DOI: 10.1007/BF00992972.

[66] R. S. McGowan and S. Nittrouer. 'Differences in fricative production between children and adults: evidence from an acoustic analysis of /sh/ and /s/'. In: *The Journal of the Acoustical Society of America* (Jan. 1988).

[67] Amina Memon and Aldert Vrij. *Psychology and Law: Truthfulness, Accuracy and Credibility Second Edition*. en. 2003.

[68] Microsoft. *Language Understanding (LUIS)*. 2021.

[69] Microsoft. *Language Understanding pricing*. English. 2021.

[70] Maria D Molina et al. '"Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content'. en. In: *American Behavioral Scientist* (Oct. 2019), p. 33.

[71] Raymond S Nickerson. 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises'. en. In: *Review of General Psychology* 2 (1998), pp. 175–220. DOI: 10.1037/1089-2680.2.2.175.

[72] Kyo-Joong Oh et al. 'A Chatbot for Psychiatric Counseling in Mental Healthcare Service Based on Emotional Dialogue Analysis and Sentence Generation'. en. In: *2017 18th IEEE International Conference on Mobile Data Management (MDM)*. Daejeon, South Korea: IEEE, May 2017, pp. 371–375. ISBN: 978-1-5386-3932-0. DOI: 10.1109/MDM.2017.64.

[73] Abby Ohlheiser and Karen Hao. *An AI is training counselors to deal with teens in crisis*. English. Feb. 2021.

[74] Christopher Olah. *Understanding LSTM Networks*. English. Github Blog. Aug. 2015.

[75] Matthew E. Peters et al. 'Deep contextualized word representations'. en. In: *arXiv:1802.05365 [cs]* (Mar. 2018). arXiv: 1802.05365.

[76] Emma Pierson et al. 'A large-scale analysis of racial disparities in police stops across the United States'. en. In: *arXiv:1706.05678 [stat]* (June 2017). arXiv: 1706.05678.

[77] Francesco Pompedda, Angelo Zappalà and Pekka Santtila. 'Simulations of child sexual abuse interviews using avatars paired with feedback improves interview quality'. en. In: *Psychology, Crime & Law* 21.1 (Jan. 2015), pp. 28–52. ISSN: 1068-316X, 1477-2744. DOI: 10.1080/1068316X.2014.915323.

[78] Debra Ann Poole and Michael E. Lamb. *Investigative interviews of children: a guide for helping professionals*. 1st ed. Washington, DC: American Psychological Association, 1998. ISBN: 978-1-55798-500-2.

[79] Martine B Powell, Belinda Guadagno and Mairi Benson. 'Improving child investigative interviewer performance through computer-based learning activities'. In: *Policing and Society* 26 (2016), pp. 365–374.

[80] Martine B. Powell and Sonja P. Brubacher. 'The origin, experimental basis, and application of the standard interview method: An information-gathering framework'. en. In: *Australian Psychologist* 55.6 (Dec. 2020), pp. 645–659. ISSN: 0005-0067, 1742-9544. DOI: 10.1111/ap.12468.

[81] Alec Radford et al. 'Language Models are Unsupervised Multitask Learners'. en. In: (2018), p. 24.

[82] Allison D Redlich and Christian A Meissner. 'Techniques and controversies in the interrogation of suspects: The artful practice versus the scientific study'. en. In: (Jan. 2009), p. 16.

[83] Radim Řehůřek and Petr Sojka. 'Software Framework for Topic Modelling with Large Corpora'. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[84] Patrick Risan, Per-Einar Binder and Rebecca Jane Milne. 'Emotional Intelligence in Police Interviews—Approach, Training and the Usefulness of the Concept'. en. In: *Journal of Forensic Psychology Practice* 16.5 (Oct. 2016), pp. 410–424. ISSN: 1522-8932, 1522-9092. DOI: 10.1080/15228932.2016.1234143.

[85] Bernardino Romera-Paredes and Philip H. S. Torr. 'An Embarrassingly Simple Approach to Zero-Shot Learning'. en. In: *Visual Attributes*. Ed. by Rogerio Schmidt Feris, Christoph Lampert and Devi Parikh. Series Title: Advances in Computer Vision and Pattern Recognition. Cham: Springer International Publishing, 2017, pp. 11–30. ISBN: 978-3-319-50075-1. DOI: 10.1007/978-3-319-50077-5_2.

[86] Peter Salovey and David J. Sluyter, eds. *Emotional development and emotional intelligence: Educational implications.* Emotional development and emotional intelligence: Educational implications. Pages: xvi, 288. New York, NY, US: Basic Books, 1997. ISBN: 0-465-09587-9 (Hardcover).

[87] Victor Sanh et al. 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter'. en. In: *arXiv:1910.01108 [cs]* (Feb. 2020). arXiv: 1910.01108.

[88] S. R. Simon, P. J. Sousa and S. E. MacBride. 'The Importance of Feedback Training'. en. In: (Jan. 1997).

[89] Richard Socher et al. 'Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank'. en. In: (2013), p. 12.

[90] Alex Wang et al. 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding'. en. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446.

[91] Helen L Westcott, Graham M Davies and Ray H.C. Bull. 'A Handbook of Psychological Research and Forensic Practice'. en. In: (2002), p. 429.

[92] Cathy Spatz Widom. 'Longterm Consequences of Child Maltreatment'. In: *Handbook of Child Maltreatment*. Ed. by Jill E. Korbin and Richard D. Krugman. Dordrecht: Springer Netherlands, 2014, pp. 225–247. ISBN: 978-94-007-7208-3. DOI: 10.1007/978-94-007-7208-3_12.

[93] Adina Williams, Nikita Nangia and Samuel R. Bowman. 'A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference'. en. In: *arXiv:1704.05426 [cs]* (Feb. 2018). arXiv: 1704.05426.

[94] Thomas Wolf et al. 'HuggingFace's Transformers: State-of-the-art Natural Language Processing'. en. In: *arXiv:1910.03771 [cs]* (July 2020). arXiv: 1910.03771.

[95] Yongqin Xian et al. 'Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly'. en. In: *arXiv:1707.00600 [cs]* (Sept. 2020). arXiv: 1707.00600.

[96] Junjie Yin et al. 'A Deep Learning Based Chatbot for Campus Psychological Therapy'. In: *arXiv:1910.06707 [cs]* (Oct. 2019). arXiv: 1910.06707.

[97] Rowan Zellers, Yonatan Bisk et al. 'SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference'. In: *CoRR* abs/1808.05326 (2018). arXiv: 1808.05326.

[98] Rowan Zellers, Ari Holtzman et al. 'Defending Against Neural Fake News'. In: *CoRR* abs/1905.12616 (2019). arXiv: 1905.12616.

[99] Yizhe Zhang et al. 'DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation'. en. In: *arXiv:1911.00536 [cs]* (May 2020). arXiv: 1911.00536.

[100] Hao Zhou et al. 'Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory'. en. In: *arXiv:1704.01074 [cs]* (May 2018). arXiv: 1704.01074.