Ariel University

Assisting Children with Special Needs in Their Daily Interaction with Other People

A research work submitted in partial fulfillment of the requirements for the degree Doctor of Philosophy by

Merav Allouche

This work was prepared under the supervision of

Prof. Vadim Levit
Prof. Amos Azaria
Dr. Rina Azoulay

June 2022

Thanks

First of all, thank G-d, who gave me the all my needs to do this work all over the years.
Next I want to thank my dear husband and the lovely children who encouraged me all the time, even when I did not have time for them.
Also to Prof. Amos Azaria that helped me and taught me a lot.
And last but not least, to Dr Rina Azoulay who is a friend and facilitator, who accompanied me step by step until the end with great patience.

Abstract

Children and adults with special needs may find it difficult to recognize danger and threats as well as socially complex situations. They are thus at risk of becoming victims of exploitation and violence. In addition, they may find themselves unintentionally insulting their friends and relatives.

The ultimate aim of this thesis is to help children with special needs better understand the environment around them and interact effectively with other people. In order to accomplish this, we propose developing an assisting agent to help people with special needs recognize risky or insulting situations. The assisting agent will detect these situations, signal the user accordingly (by text, speech, or other forms of signaling), and suggest an appropriate response. We started with text analysis and created a dataset containing 13,490 text sentences, categorized into one of four classes: "normal" sentences, insulting sentences, negative sentences about a different person, and risky sentences that may indicate a dangerous situation for a person with special needs, which requires immediate intervention. In this stage we applied several machine learning methods to 90% of the sentences randomly chosen, from the dataset, and tested them on the remaining 10%. We obtained an accuracy of close to 70% in classifying the sentences in the test set.

In the advance stage of our work we compose a text and audio dataset, which includes the text and audio of over 2600 sentences extracted from videos presenting real world situations, and categorize it into three classes: neutral sentences, insulting sentences, and risky sentences indicating unsafe conditions. We compare the ability of various machine learning methods to detect insulting and unsafe sentences. In particular, we find that a deep neural network that accepts as input the text embedding vectors of BERT and the audio embedding vectors of Wav2Vec, reaches the highest accuracy in detecting unsafe and insulting situations. Our results indicate that it may be applicable to build an automated agent that will be able to detect unsafe and unpleasant situations that children with special needs may encounter, given the dialogue contexts conducted with these children.

We have also perform a review on conversational agents (CAs). In recent years, CAs have become ubiquitous and are a presence in our daily routines. It seems that the technology has finally ripened to advance the use of CAs in various domains, including commercial, healthcare, educational, political, industrial, and personal domains. In our review, the main areas in which CAs are successful are described along with the main technologies that enable the creation of CAs. Capable of conducting ongoing communication with humans, CAs are encountered in natural language processing, deep learning, and technologies that integrate emotional aspects. The technologies used for the evaluation of CAs and publicly available datasets are outlined.

Contents

List of Figures

List of Tables

Publications

1. Utilizing Supervised and Self-Supervised Methods for Detecting Harmful Situations Based on Audio and Text, Allouch, Merav, Amos Azaria, and Rina Azoulay. submitted to IEEE Journal of Selected Topics in Signal Processing

2. Conversational Agents: Goals, Technologies, Vision and Challenges. ,Allouch, Merav, Amos Azaria, and Rina Azoulay, published at Sensors 21.24 (2021): 8448.

3. Detecting sentences that may be harmful to children with special needs, Allouch, Merav, Amos Azaria, and Rina Azoulay, published at 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2019.

4. Automatic detection of insulting sentences in conversation. Allouch, Merav, Amos Azaria, Rina Azoulay,Ester Ben-Izchak,Moti Zwilling and Ditza A. Zachor published at 2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE). IEEE, 2018.

Abbreviations

The following abbreviations are used in this manuscript:

AGATA	Automatic Ganeration of IAML from Text Acqusition
ASD	Autistic Spectrum Disorder
ASK	Alexa Skills Kit
AI	Artificial Intelligence
AIML	Artificial Intelligence Markup Language
ASR	Automatic Speech Recognition
ASRU	Automatic Speech Recognition
B2B	Business to Business
BSLTM	Bidirectional Long-Short Term Memory
CAs	Conversational Agents
CCG	Combinatory Categorial Grammar
CFG	Context Free Grammar
CharSCNN	Character to Sentence CNN
CORGI	COmmonsense ReasoninG by Instruction
CTGAN	Conditional Text Generative Adversarial Network
DAN	Deep Average Network
DBN	Dynamic Bayesian Network
DCNN	Dynamic CNN
DNN	Deep Neural Network
DSTC	Dialog State Tracking Challenge
DOAJ	Directory of Open Access Journals
DRL	Deep Reinforcement Learning
DRQN	Deep Recurrent QNetwork
DSTC	Dialog System Technology Challenge
ECA	Embodied Conversational Agent
ECM	Emotional Chatting Machine
ED	Emotion Detection
EQ	Emotional Quotient
FAQ	Frequently Asked Questions
FC	Fully Connected
GAN	Generative Adversarial Network
HMM	Hidden Markov Model
HQ	hedonic quality
HRED	Hierarchical Recurrent Rncoder-Decoder

IMDB	Internet Movie Database
IoT	Internet of Things
IQ	Itelligent Quotient
IR	Information Retrevel
IRIS	Informal Response Interactive System
IS	Information systems
ITS	Intelligent Tutoring Systems
IVR	Interactive Voice Response
JA	Joint Attention
LD	Linear Dichroism
LIA	Learning by Instruction Agent
LSA	Latent Semantic Analysis
LSTM	Long Short-term Memory
MDP	Markov Decision Process
MDPI	Multidisciplinary Digital Publishing Institute
ML	Machine Learning
MLP	Multi-Layer Perceptron
MMI	Maximum Mutual Information
MOOC	Massive Open Online Course
MR	Movie Review
MT	Machine Translation
MVRNN	Matrix-Vector Recursive Neural Network
NBT	Neural Belief Tracking
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
OAA	One against All
PCFG	Probabilistic Context Free Grammar
PBD	Programming-By-Demonstration
PoS	Part Of Speech
PRS	Procedural Reasoning System
RNN	Recurrent Neural Network
RNTN	Recursive Neural Tensor Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SAR	Socially Assistive Robotics
SAIF	Socially Aware personal assistant Implicit Feedback correction detector

SCE	Socio-Cognitive Engineering
SGD	Schema Guided Dialogue
SL	Sign Language
SQUAD	Stanford Question Answering Dataset
SSA	Sensibleness and Specificity Average
SST	Stanford Sentiment Treebank
STS	Stanford Twitter Sentiment
SVM	Support Vector Machine
TFIDF	Term Frequency Inverse Document Frequency
TFN	Tensor Fusion Network
TLA	Three Letter Acronym
TOM	Theory Of Mind
UX	User Experience

1 Introduction

Children and adults with special needs may need help in understanding their environment and interacting with others [135, 58]. They can be threatened by people, even in familiar neighborhoods around their home or school [181]. For example, such children may agree to follow strangers and are thus at high risk of harm. In addition, these children may create destructive relationships, resulting in various types of abuse and bullying towards them. In particular, they can be harmed by people with malicious intentions without even realizing it [181]. They may also speak in a way that could either be harmful to people around them, or even be used against them by being ridiculed or exploited [131].

The overall goal of our work is to develop an autonomous agent to assist children with special needs in their communication with other people. In order to help these children, the agent must be aware of the child's verbal interactions, translate the audio content into text(using ASR application), classify the text, and detect a whether an encounter requiring intervention is occurring (i.e. a risky interaction). Once the context has been established, the agent should be able to give the child feedback relevant to the situation, and even warn his/her parents or caretakers if that is warranted.

There are several types of scenarios that can challenge a child with special needs. For example, the child may unintentionally say something insulting, an unfortunately common occurrence among "special needs" children. In particular, parents often report that their child will make statements, such as "you are fat", "you are old", "go home", and "the food stinks" without realizing that they are insulting. Another concern is that a relative, or any other person, may exploit the innocence of the child to cause harm. Thus, the aim of our study is to design an assisting agent that will be able to detect insulting or risky sentences, in order to provide relevant feedback in those situations.

We had two main stages in this work: in the first stage we worked with text alone and in the second stage we combined text and voice in order to proceed towards our goal. We started by creating a dataset of approximately 13,490 sentences, that fall into the following four categories: "normal" sentences, insulting sentences[1], negative sentences about a third person, or risky sentences that may indicate a dangerous situation for people with special needs that requires immediate intervention. In this stage, we started our dataset with an initial seed of 100 unintentionally insulting sentences obtained through interviews (performed by the Autism Center[2]) with parents of children with autism spectrum disorder (ASD). To

[1]There is also a fifth category of sentences whose meaning depends on the context of the situation in which they were stated; however this category is not included in the current study because such interpretation requires additional information about the situation, rather than simply the text itself

[2]The Autism Center is part of the Department of Communication Disorders, in Ariel University

this initial dataset, we added additional sentences in all four categories from varied sources, including forums and article comments, with a focus on sentences that can be said by children, or to a child.

Next, we proceeded by selecting for evaluation several machine learning methods that could be applied to the sentence categorization task. We show that given a sufficiently large dataset of classified sentences, we are able to predict the classification of new unseen sentences, with a mean of 70% accuracy per sentence, with either the random forest or convolutional neural network (CNN) based method. We evaluated 10 CNN-based systems, each built by training with 90% of the training set and validated by the remaining 10% (the validation set). Then, we developed a panel based on the best five CNN systems, and for each sentence in the test set, a voting criterion was used for classification. Using this voting panel, the average accuracy on the test set increased to 72.2% (std=0.009, for 50 trials) and the F1 score (The F1 score is a single number that allows us to compare systems by combining recall and precision through this formula 2*(precision*recall)/(precision+recall)) to 0.714 (std=0.009), higher than all of the other individual methods.

In the advanced stage, we utilize machine learning and deep learning methods to detect insulting or unsafe situations, through the text and audio of speech. For this purpose, we have collected a text and audio database, which includes the text and audio of over 2600 Hebrew sentences extracted from videos presenting real world situations. The sentences were categorized into three classes: neutral sentences, insulting sentences, and sentences indicating unsafe conditions. We use machine learning and deep learning methods, to detect unsafe and insulting conditions using text and audio contents of these sentences. We compare the performance of different machine learning methods, and in particular, suggest using deep learning applied on text embedding using BERT [78, 29], and audio embedding using Wav2Vec [17], to detect unsafe and insulting situations. We also found that the information extracted from the spoken text is more important for detecting unsafe and insulting sentences than the information extracted from the audio only. However, the audio signals have added additional value, i.e., a system trained on both text and audio signals achieves a higher accuracy level than a system trained only on text.

Our results indicate that it may be applicable to build an automated agent that will be able to detect unsafe and unpleasant situations that children with special needs may encounter, using the dialogue contexts conducted with these children. This agent may first convert the audio input to text, and then, it may use text embedding, in addition to wav embedding, as input for a neural network, to determine the harmful situations. The ability to successfully detect unsafe and insulting context using embedded text and embedded audio indicates the applicability of building such systems to assist and protect children with special needs that

may encounter such challenging situations.

2 Related Work

In this section we have detailed works related to our topic in three areas: applications and robots in the field of special children assistance, analysis of emotions from text and analysis of emotions from text and voice.

2.1 Social Agent and Robots for Children with Special Needs

In order to help children and adults with ASD that may encounter difficulties in communicating with other people and in understanding social situations, several information communication technology-based methods were developed, in recent years. Boucenna et al. [51] provide a comprehensive review on technologies, algorithms, interfaces and sensors that can sense the children's behavior, train and improve their social abilities and train individuals to recognize facial emotions, emotional gestures and emotional situations. They suggest the use of robots to provide feedback and encouragement during skill learning interventions, and emphasize that a child with ASD might find it easier to interact with a robot than with a human teacher. The robot can provide instructions to the child who is interacting with a human therapist and encourage the child to proceed with the interaction. While their research was theoretical, in our research, we plan to develop an artificial assisting agent, that can understand the current social situation, from the words spoken by or to a child with ASD. In order to do this, we need to solve the relevant algorithmic challenges; this will enable, the assisting agent to recognize problematic situations that the child encounters.

2.1.1 Sereous Games and AI Technologies to Promote Children with Special Needs

One approach for supporting and promoting children with ASD is done by using serious games. A review by Serretc et al. [57] has mentioned 31 serious games that are used to teach social interactions to individuals with ASD. 16 of these games target emotion recognition or production and 15 target social skills. These serious games appeared promising because they can support training on many different skills and they favour interactions in diverse contexts and situations, some of which may resemble real life.

One example of serious games was developed by Jouen et al. [282], an automated platform named GOLIATH. This platform enables intensive intervention by mapping two pivotal skills in autism spectrum disorder: Imitation and Joint Attention (JA). JA introduces a third

partner during interaction .i.e. viewing the behavior of others as intentionally driven. The GOLIATH platform includes eleven games: seven Imitations and four JA. The games involved application of visual and audio stimuli with multiple difficulty levels and a wide variety of tasks and actions pertaining to the Imitation and JA.

Another important review in this field was done by Ismail et al. [129]. They examine published works that address the research gaps found in the field of robot interaction with children with ASD. They identified three major research gaps in that area: (1) Not enough diversity in research focus, (2) bias contribution in robotics research towards specific behavior defects in autism and, (3) the effectiveness of human–robot interaction after robot-based intervention duration.

Recent research on technology-facilitated diagnosis and treatment of children and adults with ASD was reviewed by Liu et al. [172]. They focus on the engineering perspective of autism studies and outlined three major delivery types of technology-facilitated autism studies: (1)Computers, Game Consoles and Mobile Devices, (2)Virtual Reality Systems/Devices, and (3)Social Robots.

Children with ASD are more comfortable connecting with social robots than with humans, which is one of the reasons this research area was developed while working with these children [275]. Tennyson et al. [275] described robotic platforms developed and investigated as a possible tool to improve social interactions among individuals with ASD.

The humanoid robot Kaspar which was developed in 2005, is also another example of assistive robotics for children with ASD. Wood et al. [298] show the development of Kaspar's design and explain the rationale behind each change to the platform. They cover the different generations of Kaspar robot and how the development of each generation expanded the robot's ability to play and learn with children with ASD.

Further robot research was done by Huskens et al. [35] who investigated the effectiveness of a brief robot-mediated intervention based on Lego therapy on improving collaborative behaviors between children with ASD and their siblings during play sessions in a therapeutic setting.

The challenge of listening to the user and understanding the user's emotional feelings is considered in Sarder's [1] thesis work, which studies the issue of conversational agent development for mental health intervention. Sarder builds an embodied conversational agent with three different levels of backchannel strategies and runs a within-subject study with a convenience sample of 24 participants. He shows that the emotional content recognized in the words of the user increases as the CA listening capabilities increase.

Our work is similar to that agent, because we need to analyze the voice and understand the relevant emotion; however, we must also mediate it to the child with special needs.

Begoli [84] presents an architecture in support of intelligent agent-mediated, behavioral interventions in special education programs for individuals with ASD. They propose a derivative of the Procedural Reasoning System (PRS) architecture.PRS is with representative, interpretative, reasoning,knowledge-based, and procedural control components abstracted from the physical aspects of the agent's placement in the environment.

2.1.2 Special Needs Education and Assistance CA

In recent years, researchers have expressed a growing interest in using CAs as well as social robots as a positive intervention for children with special needs [162].

Many disabilities fall under the heading of special needs, including: Autism Spectrum Disorder (ASD), Down Syndrome, Asperger syndrome, Rett syndrome and other sort etc , all can affect both children and adults. ASD, is a lifelong neuro-developmental disorder characterized by impaired reciprocal social communication and a pattern of restricted, often non-adaptive repetitive behaviors, interests and activities [24]. One of the widely accepted cognitive explanations for these symptoms in people with ASD is a deficit in Theory Of Mind (ToM). ToM refers to the ability of individuals to impute mental states, such as emotions, beliefs and ideas to oneself and others, and to predict the behavior of others on the basis of their mental states [40, 114]. ToM task performance is a crucial capacity which enables one to decode and understand social cues [58]. Difficulties in performing ToM tasks can impair social interactions including deficits in pragmatic abilities and empathy [19]. These deficits might lead a person to make insulting statements unwittingly, or to be unaware of verbal bulling. A high prevalence of bullying toward children and adults with ASD has been documented, including verbal bullying such as name calling and teasing (summarized in [59]).

PunkBuddy is a tool that includes a chatbot that helps dyslexic students learn through interaction. The chatbot can advise students on the rules of using punctuation, utilizing the benefits of explicit instruction [280].

Park et al. [217] developed a voice based virtual agent for children with ADHD to help them in their daily tasks. The agent provides vocal feedback to the child and encourages the child to complete the task (on time). The child reports back to the agent about her/his progress.

Xuan et al. [162] developed a chatbot dedicated to children with autistic spectrum disorder (ASD) to improve their conversation abilities. Their chatbot is intended to arouse the curiosity of children and assist them in understanding the conversation better. The chatbot uses a large question-and-answer corpus. Social assistance CAs are commonly used to assist children and adults with special needs, and especially children with ASD.

Indeed, several studies have shown that social robots can help improve social skills of

children with ASD [283], and some have indicated that a child with ASD might find it easier to interact with a social robot than with a human teacher [52].

Scassellati et al. [23] developed a social robot to increase the social communication skills of children with ASD. The robot can move or talk according to a selected task defined by the caregiver. For example, the robot can present a social situation, and ask the child what the story character is feeling. They reported that after a one-month deployment, the children with ASD improved their behavior and gained their independence.

Costa et al. [86] introduced QTrobot, a social robot developed to assist children with ASD to focus their attention, imitate positive behavior, and reduce repetitive and stereotyped behaviors. QTrobot converses with the child and plays imitation games with the child. Costa et al. show that children pay more attention to QTrobot than to a person, imitate the robot as if it is a person, and practice fewer repetitive and stereotyped behaviors with the robot than with the person.

Vanderborght et al. [36] developed Probo, which is a social story telling robot capable of expressing emotions via facial expressions and gaze. Probo uses stories to teach children with ASD how to react in different situations, such as saying "hello" or "thank you". Probo also teaches children to share their toys. Vanderborght et al. show that there are situations where the social performance of autistic children improves when using Probo.

Another known robot developed in the same project is Nao. [221], an embedded CA that has been tested and deployed in several healthcare scenarios, including care homes and schools.

Our social agent will accompany the child with special needs in his or her daily social interactions and will advise the child on proper behavior. It is different from the other agents since they are used by the child only at specific times. It also differs from the other agents because our agent must put itself in the place of the child.

2.2 Technologies Behind Emotion Recognition

In order to help a child with special needs understand the environment in which he finds himself we need to understand what emotions he perceives from the environment, and what emotions the environment perceives from him. Therefore we want to understand emotions through voice and text.

2.2.1 Text Emotion Recognition

Emotion recognition or sentiment analysis, of text and speech is often used in order to determine the sentiments and emotions of the writer or speaker [238]. As demonstrated by the

research mentioned below, over the years, a wide range of algorithms have been employed, which include both supervised and unsupervised methods. In the supervised setting, early papers used all types of supervised machine learning methods (such as Support Vector Machines (SVM), Maximum Entropy, Naïve Bayes, etc.) and a variety of feature combinations. Unsupervised methods include various methods that exploit sentiment lexicons, grammatical analysis, and syntactic patterns. Deep learning has emerged as a powerful machine learning technique and produced state-of-the-art results in many application domains as well as sentiment analysis. In our research [187], we concentrated on the insulting sentences recognition using text contents. We generated a dataset consisting of insulting and non-insulting sentences and compared the ability of different classical ML methods in detecting the insulting content.

Zhang et al. [177] reviewed current algorithms in sentiment analysis and opinion mining using deep learning. In our examination of his review we concentrate our attention on algorithms used for sentiment analysis at the sentence level and introduce part of these algorithms and methods below.

Socher et al. [247] first proposed a semi-supervised Recursive Autoencoder Network (RAE) for sentence level sentiment classification, which obtains a reduced dimensional vector representation of a sentence. Later, Socher et al. [246] proposed a Matrix-Vector Recursive Neural Network (MVRNN), an approach that builds representations of multi-word units from single word vector representations to form a linear combination of the single word representation; In Socher et al. [264], the authors introduced the Recursive Neural Tensor Network (RNTN) for the sentiment classification task.

For semantic modelling of sentences, Kalchbrenner et al. [204] propose a Dynamic CNN (DCNN), a network that uses dynamic k-Max pooling, a global pooling operation over linear sequences.

The Character to Sentence CNN (CharSCNN) model, proposed by Dos Santos and Gatti [250], proposed a Character to Sentence (CharS) model. CharS uses two convolutional layers to extract relevant features from words and sentences of any size to perform sentiment analysis of short texts.

Another approach, that achieved similar results the use of a linguistically regularized Long Short-Term Memory (LSTM) network, was presented by Qian et al. [227]. Their proposed model incorporates linguistic resources, such as sentiment lexicon, negation words and intensity words, into the LSTM in order to more accurately capture the sentiment effect in sentences.

Wang et al. [292] also utilized LSTM for Twitter sentiment classification by simulating the interactions of words during the composition process. In another study, Wang et al.

[137] combined the two previous methods and proposed a regional CNN-LSTM model, which consists of two parts: a regional CNN and an LSTM network, to predict the valence arousal ratings of text. However, the results from this method were not as strong as those of the previous methods.

Guggilla et al.[111] presented an LSTM based deep neural network model, which utilizes word2 vec and linguistic embeddings for claim classification (classifying sentences as factual or emotional).

With the goal of enhancing phrase and sentence representation, Huang et al. [191] proposed to encode the syntactic knowledge (e.g., part-of-speech tags)in a tree structured LSTM to improve phrase sentiment classification.

By combining deep learning and classical feature-based models using a Multi-Layer Perceptron (MLP) network for financial sentiment analysis, Akhtar et al. [16] employed several ensemble models for fine-grained sentiment classification of financial microblogs and news. This approach achieved slightly better results than those of Guggilla, who used a dataset with similar properties.

Guan et al.'s [110] goal was to identify each sentence's semantic orientation (e.g. positive or negative) of a review. They proposed a weakly-supervised CNN for both sentence and aspect level sentiment classification. In the first stages of our research, we combine several deep learning and context-sensitive lexicon-based methods. Teng et al. [274] proposed a context-sensitive lexicon-based method for sentiment classification based on a simple weighted-sum model, using bidirectional LSTM to learn the sentiment strength, intensification and negation of lexicon sentiments in composing the sentiment value of a sentence.

Abbasi et al. [2] used sentiment analysis methodologies for the classification of Web forum opinions in multiple languages. To achieve this goal, they considered a wide array of stylistic attributes, including lexical, structural, and word style markers, in addition to syntactic features, and they used a hybridized genetic algorithm that incorporates the information-gain heuristic for feature selection.

Another approach was used by Shaheen et al. [248], who propose a framework for emotion classification in sentences where emotions are treated as generalized concepts extracted from the sentences. They built an emotion seed that they call an Emotion Recognition Rule(ERR) and used a suite of classifiers to compare the generated ERR.

While our proposed artificial assisting agent relies generally on the ability to perform sentiment analysis, it also relies on the ability to detect hate speech, bullying, and insulting speech from two perspectives. From the point of view of the children, we would like to detect bullying directed at children with special needs, in order to protect them. From the viewpoint of those that interact with these children, we would like to see the child behave appropriately

and refrain from speaking in an insulting manner. The insulting sentences in our domain can be the result of innocent intentions and, in most cases, they do not contain language that is considered bullying behavior. Some prior work addresses that issue, Nobata et al. [206] used a Vowpal Wabbit's regression model and NLP features to detect hate speech in online user comments from two domains which outperforms a state-of the-art deep learning approach.

An approach with some similarities was used by Libeskind et al. [165]. They detect abusive Hebrew texts in comments on Facebook, using highly sparse n-gram representation of letters. Since comments in social media are usually short, they suggest four dimension reduction methods that classify similar words into groups, and they show that the character n-gram representations outperform all the other representations.

Dadvar et al. [72] propose integrating expert knowledge into the system for cyberbullying detection. Using a multi-criteria evaluation system, they obtain a better understanding of YouTube users' bulling behavior and their characteristics through expert knowledge. Based on that knowledge, the system assigns a score to users, which represents their level of bullying based on the history of their activities.

Our work is different from typical sentiment analysis, where the emotions of the writer are detected; instead, we focus on detecting the sentences that cause the listener to feel insulted or bullied. This will allow us to guide the child toward more appropriate behavior in the future; for example, choosing not to tell grandma that she is fat. Two previous studies addressed this goal Kai et al, [99] used a rule-based system underlying the conditions that trigger emotions based on an emotional model. The data set was comprised of text from Chinese micro-blogs. They used the ECOCC emotion model and extracted the corresponding cause components in fine-grained emotions.

Gui et al. [112] addressed the issue of emotion cause extraction, extracting the stimuli, or cause of an emotion. Then, they proposed an event-driven emotion cause extraction method in which a 7-tuple representation of events used. Based on this structured representation of events and the inclusion of lexical features, they designed a Convolutional kernel-based learning method to identify emotion cause events using syntactic structures.

An agent with a different goal was developed by Chkroun and Azaria [189, 188].They developed Safebot, a chatbot system that converses with humans. This system allows humans to teach it how to reply to new statements (this is similar to [31, 67]). Safebot uses human feedback to identify offensive behavior. When Safebot is told that it said something offensive, it apologizes and adds the offensive sentence to its database. It then avoids using such sentences again. There has also been work on deceptive speech detection [106, 32].

Remark: IMDB DS was extracted from MR DS.

The performance of the related work we surveyed is summarized in table 1. We can see

Table 1: Comparison of Methods for text Sentiment Classification

Research Work	**#classes**	**DB**	**Acc**
Kalchbrenner et al.[204]	2	MR	86.8
Dos Santos et al.[250]	2	MR	85.7
Wang et al.[291]	2	MR	82.28
Socher et al.[246]	2	IMDB	79
Huang et al.[191]	2	MR	89.6
Qian et al.[227]	2	MR	82.1
Abbasi et al.[2]	2	MR	91.7
Teng et al.[274]	2	MR	86.22
Yu et al.[310]	2	MR	84.9
Socher et al.[264]	2	SST	80.7
Huang et al.[191]	5	SST	52.6
Wang et al.[291]	5	SST1	50.68
Wang et al.[291]	2	SST2	89.95
Wang et al.[292]	2	SST	83
Qian et al.[227]	5	SST	48.6
Wang et al.[137]	2	SST	77.8
Socher et al.[247]	5	31,000 confessions	50.1
Wang et al.[137]	9	CVAT	55.3
Dos Santos et al.[250]	5	Twitter posts	48.3
Graves et al.[108]	2	TIMIT	78.6
Guggilla et al.[111]	2	Forum Posts	83.64
	3	User Comments	75.48
Akhtar et al.[16]	3	SemEval 2017	76.5
Guan et al.[110]	2	Amazon customer reviews	87.7
Zhao et al.[319]	2	STS OMD	82.6
Dadvar et al.[72]	2	You Tube	71
Nobata et al.[206]	2	Yahoo! Finance and News	78.1
Shaheen et al.[248]	6	Twitter posts	46.9
Ours- Allouche et all	4	Unique	70

that as the number of categories increases, the accuracy level that can be reached decreases, as it becomes more difficult to determine the correct category. This is even more relevant in situations where the category numbers are not scaled (as in SST1), but each number has a different meaning, similar to our work.

Acheampong et al. [5] survey models, concepts, and approaches for text-based ED, and list the important datasets available for text-based ED. In addition, they discuss recent ED studies, their results, and their limitations.

In a related study, Schlesinger et al. [252] focus on race-talk and hate speech. They describe technologies, theories, and experiences that enable the CA to handle race-talk, and examine the generative connections between race, technology, conversation, and CAs. Drawing together technological-social interactions involved in race-talk and hate speech, they point out the need of developing generative solutions focusing on this issue.

Chen et al. [66] proposed a conditional text generative adversarial network (CTGAN), in which an emotion label is adopted as an input channel to specify the output text. To match the generated text data to the real scene, they design an automated word-level replacement strategy such that after generating initial texts by CTGAN, they extract keywords from the training texts and replace them in the generated texts.

XiaoIce is a popular social CA, developed in 2014 by Microsoft. Zhou et al. [321] describe the design of XiaoIce as an AI companion with an emotional connection. The XiaoIce design includes intelligence quotient (IQ), emotional quotient (EQ), and a culturally sensitive personality. The IQ capacity is achieved by knowledge and memory modeling. The EQ capacity includes two key components: empathy and social skills. Both IQ and EQ are combined in a unique personality. The CA personality is defined as the characteristic set of behaviors, cognition, and emotional patterns that form an individual's distinctive character. XiaoIce's developers have designed different personas for XiaoIce to suit the preferences and desires of users in different cultures and regions. By analyzing the XiaoIce online logs, Zhou et al. show that XiaoIce understands user intent, recognizes human feelings, generates appropriate responses, and is capable of establishing a long-term relationship.

Asghar et al. [26] propose three ways to incorporate emotional aspects into encoder-decoder neural conversation models: affective word embeddings, augmenting affective objectives in the loss function, and incorporating a search for affective responses during text decoding. Affective word embedding, in 3D space, can be performed using a cognitive engineering affective dictionary. Affective objectives can be augmented in the cross-entropy loss function to generate additional emotional responses. Finally, the CA can be guided to search for effective responses during decoding. Asghar et al. show that incorporating these emotional aspects improves the quality of the CA responses in terms of syntactic coherence,

naturalness, and emotional appropriateness.

Zhou et al. [320] explain the range of challenges that exist in addressing the emotion factor in large scale conversation generation. These include: (1) the difficulty of obtaining high quality emotion labeled data since emotion annotation is a subjective task, (2) the need to balance grammar and emotion in expressions, (3) the challenge of embedding emotion information. To express emotion naturally and coherently in a sentence, they design a seq2seq generation model equipped with new mechanisms for emotion expression generation.

2.2.2 Voice Emotion Recognition

In our research, our main mission is to help children with special needs understand their environment. To do this, our agent must analyze what was said to, and by the child. Speech is the fastest and most natural mode of communication between humans. This has motivated researchers to think of speech as an efficient method of human– machine interaction. Speech emotion recognition remains challenging,and the task of extracting effective emotional features has still not been solved [30].

Recent studies on emotion recognition and hate speech detection use deep learning methods trained on audio corpora. Han et al. [116] proposed using deep neural networks (DNNs) to extract high level features from raw data as a solution to the speech emotion recognition problem. They produce an emotion state probability distribution for each speech segment using DNNs.

Nwe et al. [210] proposed a method that represents the speech signals and a discrete hidden Markov model (HMM) as the classifier by using short time log frequency power coefficients (LFPC). Performance of the LFPC feature parameters is compared with that of the linear prediction Cepstral coefficients (LPCC) and mel-frequency Cepstral coefficients (MFCC) feature parameters commonly used in speech recognition systems.Results show that the proposed system yields an average accuracy of 78%

Other researchers recently combined those two methods DNN and HMM using acoustic models that achieved good speech recognition results over Gaussian mixture model based HMMs. Li et al. [160] investigate DNN-HMMs with restricted Boltzmann Machine (RBM) based unsupervised pre-training, and DNN-HMMs with discriminative pre-training and reached an accuracy of up to 77.92%

A different type of neural network was used by Wu et al. [299]. They explored spectrogram-based representations for speech emotion classification from the USC-IEMOCAP dataset. They experimented with features from both the speech spectrogram, and from the glottal volume velocity spectrogram. Their experiments investigated whether classification performance can be improved by filtering out unwanted factors of variation such as speaker identity

and verbal content (phonemes) from speech.

Zadeh et al. [312] introduced a novel model,Tensor Fusion Network(TFN), that learns both intramodal dynamics and intermodal dynamics end-to-end. The intermodal dynamic is shaped using a fusion approach, called Tensor Fusion, which explicitly accumulates unimodel, bimodal, and trimodal interactions. The intermodal dynamics are modeled through three subnetworks which embed models, for languages, visual and acoustic, respectively.They reached to 69.4% of accuracy for binary classification.

Another method, introduced by Deng et al. [76], is an Adaptive Denoising Autoencoder that uses prior knowledge learned from a target set to regularize the training in a source group. It is based on an unsupervised field adjustment method. Their goal is to achieve a compatible feature space representation for the target sets and source and at the same time ensure the transfer of knowledge in the target field. They reached accuracy of 62.5%.

Vocal speech classification can be done by classical ML methods. Noroozi et al. [87] propose using random forests for vocal emotion recognition. This technique adopts random forests to represent the speech signals, along with the decision-trees approach, to classify them into different categories. The emotions are broadly categorized into six groups. The surrey audio visual expressed emotion database is used. The proposed method has an average recognition rate of 66.28%.

Jain et al. [132] used a Support Vector Machine (SVM) to classify the speech taken as one of the four emotions (sadness, anger, fear and happiness). They classify these emotional states with a Support Vector Machine classifier using two strategies: One against All (OAA) and Gender Dependent Classification.They reached 85.085% accuracy with the MFCC algorithm.

Our final goal is online emotion recognition in order to mediate the environment for children with special needs. There has been research to develop online speech emotion recognition systems: Bertero et al. [44] show an interactive dialogue system to recognize user emotion and sentiment in real time. These conventional dialogue systems were built based on modules that enable them to have "empathy" and answer to the user while being aware of their emotion and intent. They used a CNN model to extract emotion from raw speech input without feature engineering. This approach has achieved an accuracy of 65.7%.

Nivasch and Azaria [22] introduced an architecture called Socially Aware personal assistant Implicit Feedback correction detector (SAIF). SAIF obtains pairs of two voice commands including both the user's voice and the commands' transcripts and classifies the pair of sentences either as two new commands, or as the latter command being a correction of the previous command.

In a recent study, Zhu et al. [308] suggest how to combine BERT text embedded vectors with Wav2Vec audio embedded vectors for the task of Dementia Detection. They use the

Wav2Vec model to generate Automatic Speech Recognition (ASR) transcripts and use the vectors to fine-tune BERT, followed by inference layers consisting of a Convolutional layer, a global average pooling layer, and a fully connected layer for the dementia detection task. In our study, we utilize both the Wav2Vec and Hebrew BERT embedded vectors, for the aim of risky and insulting sentences detection. In particular, we used the Wav2Vec model for fine-tuning and extracting audio features from our data, then we combined the given embedded vector with the text HebBERT embedded vector, and this concatenate vector is used to train our DNN model.

In our work we will need to recognize a wider range of emotions, and to identify when the child's behavior is "strange". We also need to give appropriate behavioral advice to the child with special needs in each situation. For example, an alert signal will be sent when bullying is recognized. We combined both text and voice and reached up to 80% accuracy.

3 A Comprehensive Review of Conversational Agents

Our research is concerned with developing an agent to assist special children in their social interactions. As part of the study, we conducted a comprehensive review of the topic of call agent development, in different areas, and using different technological tools. In this chapter we would like to describe the results of the review on

Conversational agents (CA) are agents that interact with users via written or spoken natural language. CAs accept as input natural language as speech, text, or video; in addition, they may receive input from several different sensors. CAs are required to process the input and provide relevant advice or feedback in the form of text, speech, or by manipulating a physical or a virtual body. Some CAs are capable of taking specific actions either in the real world, or in the virtual world. Most CAs use natural language processing to understand and generate speech and some may also have engagement and personalization abilities. The rapidly growing abilities introduced by modern machine learning techniques facilitate the development of CAs capable of carrying out meaningful conversations with humans, learning to generate better and more relevant responses, expanding their knowledge-base, and performing actions beneficial to their users.

Current technological development enables the increasing use of CAs in several domains, such as assistance agents in the educational domain and health system, customer support agents in the commercial domain, and influence bots in the political domain. Commercial CAs for personal use, such as Siri [50] of Apple, Meena [9] of Google, and Cortana [45] of Microsoft, are widely used around the world. The aim of our work is to outline the principles behind the development of CAs, and to survey the main domains in which conversational

agents are successfully used. Several recent studies have been carried out over the last years on CAs, and in particular, on text-based CAs that are called chatbots (as defined in Section 3.1). Some studies concentrate on the technologies behind the development of CAs, and other studies examine their impact on people, i.e., the way people interact with them and perceive them.

Several recent reviews survey CA development and usage, at times referring to them as chatbots. Adamopoulou and Moussiades [6] provide a historical perspective of the chatbot development process, present a complete chatbot categorization system, and analyze the two main approaches in chatbot development: pattern matching and machine learning. They mention two limitations of the current generation chatbots in understanding and producing natural speech, and they also point out that today's technology aims to build chatbots that can learn to talk but cannot learn to think.

In another study, Adamopoulou and Moussiades [7] present an overview of the evolution of the international community's interest in chatbots, discuss the motivations that drive the use of chatbots and their usefulness in a variety of areas. They clarify the technological concepts and classify them based on various criteria, such as the area of knowledge and the need they serve. Furthermore, they present the general architecture of modern chatbots while also mentioning the main platforms they were created for. In another study, Nuruzzaman et al. [208] present a survey on commonly used chatbots and the underlying techniques. They focus on response generating chatbots. In this category, the various response models can be categorized into four groups: template-based, generative, retrieval-based, and search engines. They compare 11 most popular chatbot application systems and present the similarities, differences, and limitations. They conclude that despite recent technological advances, chatbots conversing in a human-like manner are still hard to achieve.

Another survey concentrating on the technologies used by CAs is that of Borah et al. [49]. They describe the overall architecture of CAs, concentrating on the machine learning layer, and analyze the recent development of text-based CAs. Chen et al. [65] describe the technology behind CAs and dialog systems in real world applications and discuss the effect of recent advances in deep learning on CA development. They emphasize that "big data" available from conversations on social media can be useful in building data driven, open domain CAs capable of responding to nearly any query. They further state that deep learning technologies can be used to leverage the massive amount of data to advance CAs from different perspectives. Gao et al. [133] concentrate on deep learning based CAs. They group the conversational agents into three categories: question answering agents, task-oriented dialogue agents, and chatbots. For each category, they present a review of state-of-the-art neural approaches, draw the connection between neural and traditional approaches, and discuss the

progress that has been made and challenges still being faced using specific systems and models as case studies.

Diederich et al. [80] review 36 studies on CAs in information systems (IS). They classify the literature along five dimensions. Three dimensions are related to CAs: mode of communication, context, and embodiment; and the other two dimensions are related to IS: theory type and research method. Wolff et al. [190] define a set of criteria to categorize chatbot applications. They review 52 articles describing chatbots. Most of the papers focus on customer support chatbots, e.g., chatbots used to acquire information on specific services or products.

3.1 Related Definitions and Terms

Conversational agents are highly referenced in the literature by numerous sources, including research papers, industry documentations, and internet blogs. Unfortunately, there exist inconsistencies in the references with respect to several central concepts related to conversational agents. Therefore, the aim of this section is to improve clarity, by providing definitions for the main relevant concepts currently in use, such as conversational agents, dialog systems, chatbots, and virtual assistants.

It was observed that there are two terms that are sometimes used interchangeably: the term *conversational agent*, and the term *chatbot*. There have been several attempts to define the distinction between the two terms. According to Vishnoi's definition [288], chatbots are software components that are designed to respond to human statements with a specific set of predefined replies. However, conversational agents are more contextual than chatbots and use more advanced technologies such as deep learning methods and natural language understanding (NLU).

According to Nuseibeh [209], conversational agents are all types of software programs that interpret and respond to statements made by users in natural language. Chatbots, according to this definition, are a type of CA designed to simulate conversations with human users. Other types of CAs are programs designed to perform a particular goal, such as vacation planning and booking. CAs of this type are called *goal-oriented conversational agents*.

Radziwill and Benton [230] define conversational agents as software systems that mimic interactions with real people. They define chatbots as CAs that are implemented using a text-based interface.

Hussain et al. [127] classify chatbots into two main categories: task-oriented chatbots and non-task-oriented chatbots. According to Hussain et al., task-oriented chatbots are designed to accomplish specific goals such as ordering a pizza, guiding a user on social media, etc. The non-task-oriented chatbots for entertainment converse with users in an open domain. Masche

and Le [182] categorize conversational systems into chatbots and dialog systems. According to their definition, chatbots are systems mainly based on pattern matching, while dialog systems are based on theoretically motivated techniques that enable conversations. Nimavat and Champaneria [203] distinguish between four criteria that can be used to classify chatbots: knowledge domain, type of service provided, chatbot goal, and the response generation method. They define conversational bots as bots that talk to the user like another human being, in an open domain. It is worth noting that due to the ambiguity in the related terms and definitions, and the lack of a commonly agreed upon standard on the meaning of chatbot, the Alexa prize competition, set up with the goal of furthering conversational AI, uses the term *socialbot* to describe the conversational agents. These agents are intended to interact on a range of open domain conversational topics [287].

In this review, our own definition for CA is provided, which is built upon the definitions provided in previous studies. To properly define CA, the more general concept of dialog systems is introduced first. A *dialog system* is a human-computer interaction system that uses natural language to communicate with the user.

A *conversational agent* is a dialog system that can also understand and generate natural language content, using text, voice, or hand gestures, such as sign language. Thus, to be categorized as CA, the condition is, according to our definition, being able to understand and produce *sentences* in natural language. As a result, a CA is required to handle natural language that is not limited to a predetermined set of words (e.g. only numbers or a set of keywords) or a limited sentence structure.

The following examples cannot be considered CAs: (a) An interactive voice response (IVR) system in which the user is instructed to press a number on a keypad or say a specific word in order to advance to the next menu (e.g.: "Press or Say 1 for English"), is not considered a CA, since the user response does not include natural language *sentences*. (b) An embedded system in which a user provides voice commands (e.g. "Turn on the lights.", "Set the temperature to 25 degrees.") and the system executes them without invoking any natural language response.

There are different criteria for categorizing CAs: mode of communication, action capabilities, and the domain/application in which the CA operates. First our definition of conversational agents is refined according to the mode of communication between the CA and the human user. Here, a *chatbot* is defined as a CA that interacts with the user only by text and not by any other means of communication, for example, the ELIZA chatbot [293], or chatbots available on service platforms, such as banks, booking, and other e-commerce domains. Voice based virtual agents are CAs that interact with the users by voice, for example, Siri, Google Now, Cortana, etc. Graphically embodied agents are virtual agents that have a virtual body as well as voice understanding and speech generation abilities. Their virtual body

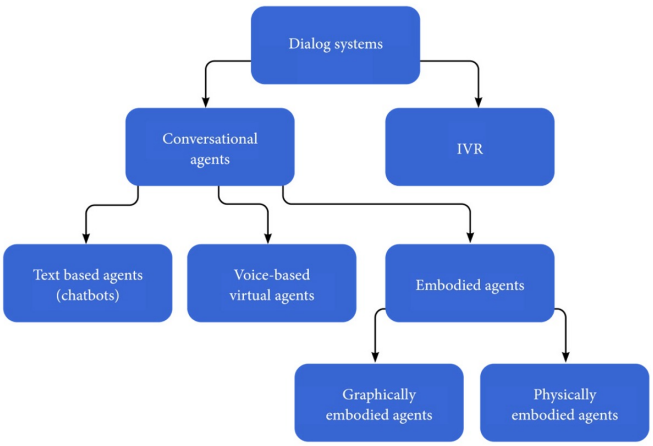

Figure 1: Conversational agents and chatbots: the definitions used in this work

enables them to provide an additional means of communication through gestures. Finally, physical-based embodied agents are CAs that have a physical body, such as social robots, e.g. JIBO [53]. Both graphical and physical agents are called embodied CAs (ECAs). The above definitions are used throughout this work and summarized in Fig. 1.

CAs can also be classified according to their effector capabilities and actions. Communication-only agents merely communicate with a user and do not execute any action e.g., ELIZA [293], Cleverbot [101, 123], or CAs used only to answer questions. Other CAs, known as virtual or personal assistants e.g., Alexa [173], are capable of executing physical or virtual actions, such as turning on an AC or booking a flight (see Fig. 2).

Finally, CAs can be classified according to the application: (a) Open domain / general purpose CAs are mainly used to answer questions in various domains or in entertainment, and are mostly communication-only agents. (b) Goal-oriented CAs assist users in completing tasks requiring multiple steps and decisions. Goal-oriented CAs are also task-oriented dialogue systems [322] and are referred to as taskbots according to the Alexa Prize competition [273]. These agents may be used both in the business domain or as personal assistants. In the business domain, they operate as customer service and sales representatives. As personal support agents, they can assist the user in particular tasks, such as driving, vacation planning, or trip management. (c) Social supporting agents can support patients in medical conditions or support students in the learning process. (d) Social network bots, also known as influence agents, are intelligent CAs acting in the social media to advertise a product or influence

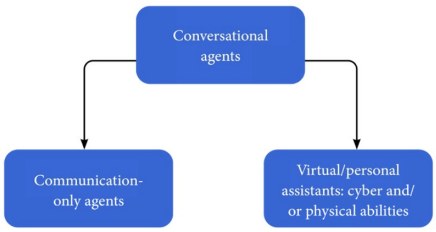

Figure 2: Conversational agent classification according to action capabilities.

opinions (see Fig. 3).

3.2 CA's Design Issues

This section describes the different components related to CA design. CA design is divided into four classes: text components for chatbots; CA components related to voice based virtual agents; physical related components for goal oriented CAs or for embodied agents; and task performance components for goal oriented CAs. For each of the four classes, the general goal is provided, the main components are detailed, and the relations between these components are described.

Text Related Components

The two main abilities required of CAs are the ability to logically understand the user's utterance and the ability to correctly reply to it. Overcoming these challenges require research in fields of natural language processing (NLP), information retrieval (IR) and machine learning (ML) [133].

Text related components are used by most CAs, including embodied CAs and voice based CAs, since voice based virtual agents usually translate human speech to text, analyze the text, generate text responses, and then produce the speech signals. Therefore, in our design description, text related components are discussed first.

CAs are commonly partitioned into components based on a pipeline determined by the order in which the component is used [94, 140]. The most common components are

- The natural language understanding (NLU) component: interprets the words into an

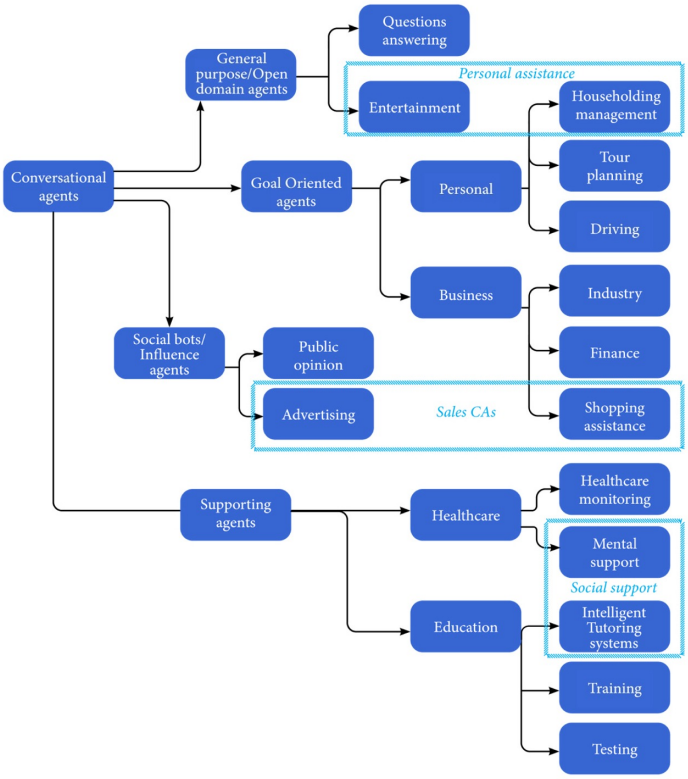

Figure 3: Conversational agent applications

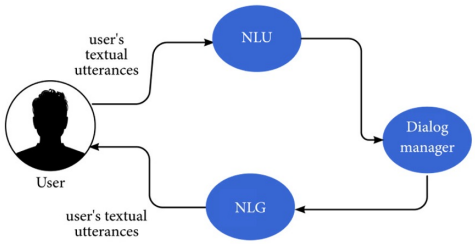

Figure 4: The textual components of CAs.

internal computer language, called a logical form, which represents the meaning of the text.

- The dialog manager component: receives the logical form and decides on how to respond. The dialog manager may also include a module that assists with long-term conversations.
- The natural language generation (NLG) component: converts the answer into a text sequence in natural human language.

A schematic description of the textual processing components is provided in Fig. 4.

Masche and Le [182] use a similar categorization, with an additional preprocessing component. They provide an alternative hierarchical approach to define text related components by dividing the components into those responsible for text understanding, text processing, and text producing, as defined by Stoner et al. [268], as follows:

- Responder - the interface between the user and the CA: transfers and monitors the inputs and the outputs.
- Classifier - the interface between the responder and the graphmaster: normalizes and filters user inputs, and processes the graphmaster output.
- Graphmaster - the brain behind the CA: manages the high-level algorithms.

According to this approach, the responder component includes parts from both NLU and NLG, while the dialog manager component has parts from both the classifier and the graphmaster.

Abdul-Kader et al. [4] survey the techniques used to design CAs, and describe the main techniques used by pattern matching based CAs, which are: (a) Parsing: manipulation of the input text using NLU functionality. (b) Pattern matching: analyzing user input and collecting relevant data, especially used by question-answering systems. (c) Chat script: used when no matches occur. (d) History database: used to enable the chatbot to remember previous conversations. (e) Markov Chain: enables probabilistic based responses of chatbots.

Ramesh et al. [234] describe various approaches to design and build chatbots. Ahmad et al. [13] provide some examples of chatbots, describe their design, and provide a description of the most popular techniques used by chatbot developers. Diederich et al [81] analyze 51 CA platforms to develop a taxonomy that would allow the identification of platform archetypes in CA design. The taxonomy consists of eleven dimensions and three archetypes, which can be used by practitioners in the design stages of CA. Lokman and Ameedeen [25] categorize modern chatbot design into the following elements: domain knowledge, response generation (retrieval or generative), text processing (vector embedding or Latin alphabet), and machine learning (ML) (mostly using neural networks). The various components described in this section enable the creation of a CAs that are able to communicate with humans through an appropriate textual interface. In the next section these technologies are also used for other types of CAs, such as voice based CAs.

Voice Related Components

Voice based virtual agents are CAs that communicate with humans using speech. The process used by CAs usually includes: translating the sound waves into text, understanding the text, producing a text response for the user, and translating the text response to the sound produced by the computer or by the robot. The steps of understanding the text and producing an answer usually rely on the text related components described above, but there are additional components, such as voice based virtual agents related to audio analysis and audio production. A voice based virtual agent may extract additional non-verbal information from the user audio, such as the user's emotional state, whether the user is being sarcastic, dramatic, decisive, or trying to deceive the system. Some works have also used non-verbal cues to detect whether a user is trying to correct previously made statements [22]. The components responsible for additional voice-based capabilities include:

- Automatic speech recognition (ASR) component (speech to text): converts the audio stream to a text representation.
- Non-verbal information extraction component: extracts relevant non-verbal informa-

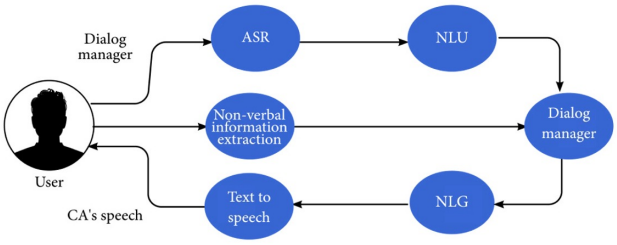

Figure 5: The main voice-based components of CAs.

tion from the audio, such as observing the user's emotional state or understanding the urgency.

- Text to speech component: synthesizes the output waveform that is sent to the speakers.

The main components of the audio process components are described in Fig. 5.

Additional information on the capabilities and components of speech-based CAs is described by Saund [251]. Benzeguiba et al. [43] review ASR challenges and technologies, and Yu and Deng [309] provide a complete overview on modern ASR technologies with an emphasis on the deep learning methods adopted in ASR.

Physical Related Components

Physical embedded CAs, which obtain visual input from the user, benefit from the ability to understand physical related gestures, such as body language and facial expressions. In addition, embodied CAs (ECAs) can use facial expressions and body gestures in their reactions.

Sign languages are complete languages that use only physical gestures to communicate. These languages may be used by CAs designed to communicate and/or tutor deaf users. Next, the main components in building an agent with these capabilities are described while referring the reader to articles reviewing this field.

Sadeghipour and Kopp [249] describe an overall model for cognitive processes of embodied perception and generation. According to them, the main components for physical agent-human communication are as follows:

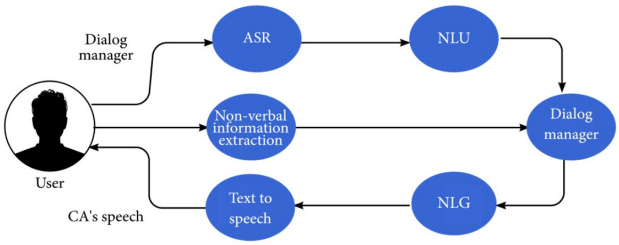

Figure 6: The main components of a physical based embodied CA.

- Perception component: receives visual movements and preprocesses them. The preprocessing pipeline consists of four submodules: (1) the body correspondence solver is responsible for performing required operations (such as rotation and scaling) on the observations. (2) the sensory memory receives the transformed positions and buffers them in chronological order. (3) the working memory holds a continuous trajectory for each hand through agent-centric space. (4) the segmenter submodule decomposes the received trajectory into movement segments called guiding strokes.

- Shared knowledge component is responsible for the representation of motor knowledge. This component consists of a hierarchical structure, starting with the form of single gesture performances in terms of movement trajectories and leading into less contextualized motor levels and then toward more context. The motor representation hierarchy consists of three levels: motor commands, motor programs and motor schemas.

- Gesture Generator component is invoked by a prior decision to express an intention through a gesture. This component may also be used by a virtual agent that is built on a motor control engine.

The main components of the physical-based embodied CA are described in Fig. 6. Krishnaswamy et al. [148]. provide a review on sign languages and gesture interpretation and generation. Homburg et al. [125] describe the process of sign language (SL) translation, including SL recognition and SL generation. Singh et al. [262] detail the process of recognizing and interpreting the Indian sign language. Finally, Beck et al. [42] study the generation of emotional body language to be displayed by humanoid robots.

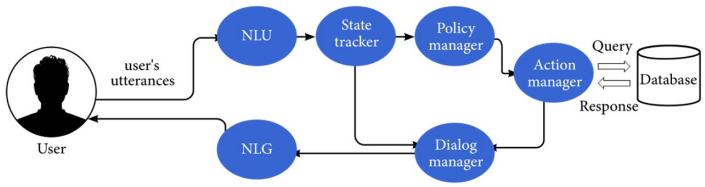

Figure 7: The main components of a goal-oriented CA.

Task Related Components

Goal oriented CAs assist users in completing tasks requiring multiple steps and decisions, such as CAs booking vacations and planning trips. Goal oriented CAs may use the text related and voice related components described above, in addition to task related components. Task related components are special components that handle task related planning and learn challenges for the successful execution of the required goal. Previous studies on goal oriented CAs [318, 87], describe the processes followed by a conventional goal-oriented CA. This process includes the phases of text understanding, state estimation, dialogue policy, and text generation. The additional task related components are defined as follows:

- State tracker: estimates the state of the user's goal by tracking the information across all turns of the dialogue.
- Policy manager: determines the next set of actions to help reach that goal. The policy manager uses the goal related information from the state tracker, and may communicate with the dialog manager.
- Action manager: performs the required cyber actions (e.g., hotel reservations, food ordering, flight booking), and/or the required physical actions to successfully fulfill the user requests.

The schematic description of the task related components is provided in Fig. 7, and an overview of the technologies behind goal oriented CAs is provided in Section 3.3.

3.3 Technologies Behind CA Components

In this section the technologies behind the CA components presented in Section 3.2 are described in further detail, detailed examples are provided for the physical components, and the

implementation of the technologies in recent CA systems are discussed.

Natural Language Understanding

Natural language understanding (NLU) typically refers to extracting structured semantic knowledge from text. NLU tasks mainly include tokenizing the text, normalizing it, recognizing the text entities and performing dependency or constituency parsing. The traditional NLU stack is based on the following five components: phonology, morphology, syntax, semantics, and reasoning [46].

In particular, morphological analysis or parsing can be viewed as resolving natural language ambiguity at different levels by mapping a natural language sentence to a series of human-defined, unambiguous, symbolic representations, such as part of speech (POS) tags, context free grammar, and first-order predicate calculus. NLU includes the following sub areas: resolution, discourse analysis, machine translation, morphological segmentation, named entity recognition, POS Tagging, and more [140]. For a review on natural language understanding, the reader is referred to the survey of Navigli [199], in which several NLU approaches and modes are reviewed, including explicit versus implicit learning, representation of words and semantics, and a vision on what machines are expected to understand.

In the remainder of this section, the focus is on studies that use NLU for CA development. Initially, CAs using classical NLU technologies are described. Next, CAs using a parser as their NLU component are described. To conclude, recent CAs that use advanced technologies for NLU are described.

A classical approach for designing chatbots is the pattern matching approach, in which the CA matches the user input with a pattern and chooses the most suitable response stored in its predefined text corpus. One example of a CA that is based solely on simple pattern matching is ELIZA [293]. Over the years, several studies have developed additional rules and corpora to develop more adaptive and advanced CAs. Inui et al. [130] use a linguistic corpus to design a CA interface. The dialogue corpus is based on a series of dialogues, and NLU is achieved by adopting corpus-based methods like the stochastic model, n-gram model, keyword matching, and structural matching.

ALICE [290] is a chatbot based on AIML [180], an XML based language designed to create chatbots based on pattern matching. ALICE won the Loebner Prize as "the most human computer" at the annual Turing Test contests of 2000, 2001, and 2004. ALICE answers the user's query by using its pattern-matching engine, which searches for a lexical correspondence between the user's query and the chatbot's patterns.

Agostaro et al. [11] outline the limitations of the pattern matching approach. Pattern

matching may fail to answer the user query when the query is composed of words that do not match any pattern. Therefore, when the query is grammatically incorrect, the pattern matching mechanism will fail. To overcome these limitations, Agostaro et al. developed LSA-bot [11], which is a chatbot based on latent semantic analysis (LSA). LSA applies statistical computations to a large corpus of text to extract and represent the meaning of words. LSA-bot uses LSA to map its knowledge base into a conceptual space. The user input is mapped into the same conceptual space, allowing LSA-bot to find an appropriate response.

The informal response interactive system (IRIS) chatbot, developed by Banchs and Li [37], uses a large database of dialogues to provide candidate responses to a given user utterance. The IRIS response selection process chooses the candidate utterances using two scores: The first score is determined by the cosine similarities between the current user input vector and all single utterances stored in the database. The second score is determined by the cosine similarity between the current vector dialogue and the dialogue history of the user. The two scores are combined using a log-linear scheme. IRIS randomly selects one of the top ranked utterances as its response.

Context free grammar (CFG) parser (e.g. [88]) is often used by CAs for NLU. A CFG parser builds a constituency parse tree from the given user utterance based on a grammar, which is composed of parsing rules. A more generalized CFG, which is more suitable for solving ambiguity, is the probabilistic CFG (PCFG) [240, 98]. In a PCFG parser, each rule in the grammar is associated with some probability. A PCFG parser outputs the parse tree with the highest probability.

Azaria et al. [33] present LIA, an agent that uses a combinatory categorial grammar (CCG) parser as its NLU component. The parser maps the commands, which are given in natural language, to logical forms, which contain functions and concepts that can later be executed by the dialog manager. CCGs benefit from being more expressive than CFGs as they can represent the long-range dependencies appearing in some sentences (e.g. relative clauses), which cannot be expressed using CFGs. Recent ML methods and word embedding methods are widely adapted to achieve NLU components with higher performance. Rasa NLU and Rasa Core [48] are open source Python libraries for building conversational software. Rasa NLU allows the use of a predefined pipeline for the NLU process. Recent ML methods and word embedding methods are widely adapted for achieving NLU components with higher performance. Rasa NLU and Rasa Core [48] are open source Python libraries for building conversational software.

Rasa NLU allows the use of a predefined pipline for the NLU process. Their recommended pipeline process starts by tokenizing the user input, followed by the conversion of each token to a GloVe embedding vector [222]. Then, a multiclass support vector machine (SVM) [70]

is used for deciding which action to take. Custom entities are recognised using a conditional random field [150].

ConvLab-2 [322], which is an open-source toolkit for building goal oriented CAs, provides three NLU models: a semantic tuple classifier, a multi-intent language understanding model [154], and a fine tuned BERT [78] based NLU model with the ability of intent classification and slot tagging.

The Dialog Manager

Given the input text, the next step in the CA's pipeline is to manage the dialogue with the user. The *dialog manager* component is responsible for two main tasks: *Dialogue modeling*: keeps track of the state of the dialogue and *Dialogue control*: decides on the next system action [186].

Harms et al. [117] review the state-of-the-art commercial and research tools available for CA dialog management. They divide the management approaches into two types: handcrafted-rule-based approaches and probabilistic (data-driven) approaches. The handcrafted dialog manager defines the state and the control of the system by a set of rules that are defined by developers and experts, while the probabilistic dialogue manager learns the rules from actual conversations.

The studies described next concentrate on dialog managers, including handcraft-rule-based systems and probabilistic-based systems. Handcraft rule-based management systems may be based on a planning algorithm or a pattern matching based approach. Nguyen and Wobcke [201] propose a planning-based approach for developing a personal assistant CA. In their approach, the dialogue manager has a set of plans, which can be divided into four groups: conversational act determination and domain task classification, intention identification, task processing, and response generation.

CommandTalk is a spoken language interface for a battlefield military simulator [196, 267]. It manages the representation of linguistic context, interprets user utterances within that context, and plans system responses. The CommandTalk dialogue manager uses a dialogue stack, a recovery mechanism for the stack, reference mechanisms, as well as finite state machines.

The MindMeld Conversational AI platform [194] is a platform designed for building conversational assistants. It uses pattern-matching rules to determine the dialogue state and based on this state and the predefined business logic, the CA performs the required task (or response) related to this state.

The Bottery CA creation platform [146] consists of four components: a set of states, a blackboard style memory, an optional set of global transitions to allow the agent to switch from state to state, and an optional grammar used by the agent to generate the final outputs of the CAs. The Bottery syntax can be simply expressed by using structured JSON and can

be extended by using imperative JavaScript code. The Bottery conversation management is performed by a finite state machine, displayed as a graph.

We proceed by describing probabilistic-based dialog management schemes. Google DialogFlow [105] is a framework for composing CAs. The Google dialog manager considers the intent or motivation extracted from the user conversation to determine the appropriate action. Another commercial CA framework is Microsoft LUIS [295], a cloud-based conversational AI service, that uses ML to understand the conversation to extract relevant information. LUIS can assist developers, who are unfamiliar with ML methods, to create their own cloud-based ML models, specific to the application domain. Herderson et al. [119] present a word-based approach to dialog state tracking using recurrent neural networks (RNNs). The model is capable of generalizing to unseen dialog states hypotheses. For long-term effects of the conversation, dialog managers consider the conversation as a Markov decision process (MDP) and choose their responses by using RL methods. Singh et al. [261] suggest using RL for goal-oriented dialog management.

Li et al. [159] suggest applying DRL to model future rewards in CAs. The agent's reward is determined according to three useful properties: informativity (non-repetitive turns), coherence, and ease of answering. The dialog manager of the ensemble-based CA developed by Serban et al. [257] for the Amazon Alexa Prize competition utilizes an ensemble of NLG and retrieval models, including template-based models, bag-of-words models, sequence-to-sequence (seq2seq) neural networks, and latent variable neural network models. Their dialog manager is trained to select an appropriate response by applying RL. The training was carried out on crowdsourced data as well as on real-world user interactions data.

Natural Language Generation

The NLG component translates the CA's representation of the response to natural language. NLG is defined by Reiter and Dale [239] as a subfield of AI and computational linguistics that is concerned with producing understandable texts in some human language from some underlying non-linguistic representation of information. Gatt and Krahmer [100] provide a recent survey on state-of-the-art NLG research, focusing on data-to-text generation. They discuss NLG architectures and approaches and highlight several new developments. In addition, they review the challenges of NLG evaluation, and show the relationships between different evaluation methods.

NLG can be performed by template-based systems, which map the non-linguistic input directly to the linguistic surface structure without intermediate representations. Van Dimter et al. [284] describe several template-based systems and compare them to other NLG systems

in terms of their potential for performing NLG tasks. They claim that template-based systems can, in principle, perform all NLG tasks in a linguistically well-founded way.

Several recent CAs use deep neural networks (DNNs) to perform the natural language generation task. Wen et al. [294] present a statistical language generator, based on a semantically controlled long short-term memory (LSTM) structure. The LSTM generator is trained on unaligned data by jointly optimizing sentence planning and surface realization. Variations in natural language output are obtained by randomly sampling the network output.

Tran et al. [279] present a semantic component, called an aggregator, which can be integrated into an existing RNN encoder-decoder architecture, to improve NLG performance. The proposed component consists of an aligner and a refiner. The aligner is a component that computes the attention over the encoded input information, while the refiner is a gating mechanism stacked over the attentive aligner to further select and aggregate the semantic elements.

Jeraska et al. [136] focus on language generation models with inputs structured for meaning representation to describe a single dialogue act with a list of key concepts that need to be conveyed to the user. They present a neural ensemble encoder-decoder model for generating natural utterances from the meaning representations.

Dusek et al. [83] assess the capabilities of recent seq2seq data-driven NLG systems, which can be trained on pairs of sequences, without the need for fine-grained semantic alignments. These pairs of sequences are composed of meaning representations, which are the output of the dialog manager and the corresponding natural language texts. They find that seq2seq NLG systems generally score high in terms of word-overlap metrics and human evaluations of naturalness, but often fail to correctly express a given meaning or representation if they lack a strong semantic control mechanism during decoding. Moreover, they can be outperformed by hand-engineered systems in terms of quality, complexity, and diversity of outputs.

End to End Models

A popular end-to-end technique used by CAs is based on sequence-to-sequence learning models. These models convert sequences from one domain into sequences in another domain. Sequence-to-sequence models are widely used in different domains, such as machine translation, text summarization, speech to text conversion, image caption generation, and automated answer generation.

Sordoni et al. [265] present a sequence-to-sequence based chatbot, trained end-to-end on large quantities of unstructured Twitter conversations. A neural network architecture is used to address sparsity issues that arise when integrating contextual information with classic

statistical models, allowing the system to take into account previous dialog utterances. They extend the recurrent neural network language model [192] and propose a set of conditional language models in which past utterances are encoded in a continuous context vector to help generate the response.

Li et al. [157] propose a method for defining the sequence-to-sequence objective function. They propose using MMI, a measurement of the mutual dependence between inputs and outputs, as the objective function for the generated conversational responses. They also present practical strategies for neural generation models that use MMI as the objective function. Experimental results demonstrate that the proposed MMI models produce more diverse, interesting, and appropriate responses, yielding substantial gains in BLEU scores and in human evaluations.

Serban et al. [256] investigate the task of building open domain CAs based on large dialogue corpora using generative models. Generative models produce responses that are generated word-by-word, opening the possibility for realistic, flexible interactions. In their model, a dialogue is considered as a sequence of utterances that, in turn, are sequences of tokens. They extend the hierarchical recurrent encoder-decoder (HRED) neural network to the dialogue domain. Their experiments demonstrate that the hierarchical recurrent neural network generative model outperforms both n-gram based models and baseline neural network models in the task of modeling utterances and speech acts. In addition, they show that the performance of their system can be improved by bootstrapping the learning from a larger question answer pair corpus and from pretrained word embeddings.

Some studies concentrate on seq2seq learning for question answering chatbots. He et al. [118] suggest a model based on sequence-to-sequence learning for a question answering chatbot, which can answer complex questions in a natural manner. The model incorporates copying and retrieving mechanisms in a bi-directional RNN. The semantic units in the answers are dynamically predicted from the vocabulary, copied from the given question, and/or retrieved from the corresponding knowledge base.

Qiu et al. [228] present a hybrid open-domain question-and-answer chatbot that combines information retrieval and seq2seq models. Information retrieval methods are used to retrieve a set of question / answer pairs based on a chat log of an online customer service. Then, the seq2seq model is used to rank the candidate answers. If the score of the top candidate answer is above a predefined threshold, it is considered to be the answer; otherwise, the answer is generated by the seq2seq model. Similarly, Ghazvininejad at el. [102] present a general data-driven and knowledge-grounded CA. They condition the CA responses not only on the conversation history but also on external facts through multi-task learning. This makes the CA versatile and applicable to an open-domain setting.

End-to-end models can also be useful in goal-oriented CA developments. Ham et al. [115] describe the use of end-to-end models for goal-oriented CAs, which need to integrate external systems to provide an explanation for the particular responses. They present an end-to-end monolithic neural model that learns to follow the core steps in the dialogue management pipeline. The model outputs all the intermediate results in the dialogue management pipeline to enable integration with the external system and to interpret why the system generates a particular response.

Kim [142] presents an end-to-end document-grounded, goal-oriented CA that utilizes a pretrained language model with an encoder-decoder structure. The encoder solves both the knowledge-seeking turn detection task and the knowledge selection task; the decoder solves the response generation task.

Das et al. [75] suggest using DRL to learn the policies of goal-oriented CAs to answer visual questions. They pose a cooperative dialogue between two CAs communicating by natural language. The dialogue involves two collaborative CAs; one CA sees the image; and the second CA asks the first one questions about the image. DRL is used for learning the policies of these agents during the multi round dialogue. As a result, the two trained CAs invent their own communication protocol without any human supervision.

Technologies Specific to Goal Oriented CAs

In the development of goal oriented CAs, there are additional challenges due to the need to combine both the dialogue handling and the task performance management. Several ML based technologies are commonly used to handle these challenges.

Zhang et al. [317] review the recent advances in goal oriented CAs and discuss three critical topics: data efficiency, multi-turn dynamics, and knowledge integration. They also review the recent progress on task-oriented dialog evaluation and widely used corpora, and they conclude by discussing some future trends for task-oriented CAs.

Zhao and Eskenazi [318] discuss the limitations of the conventional goal-oriented CA pipeline and suggest an alternative end-to-end task-oriented dialog management framework. In their framework, the state tracker is an LSTM-based classifier that inputs a dialog history and predicts the slot-value of the latest question. The policy manager is implemented by a deep recurrent Q-network (DRQN) that controls the next verbal action. This framework enables the creation of a CA, which can interface with a relational database and learn policies for both language understanding and dialog strategies.

Noroozi et al. [207] present fast schema guided tracker (FastSGT), which is a BERT-based model for state tracking in goal oriented CAs. FastSGT enables switching between

services and accepting the values offered by the system during the dialogue. Finally, an attention-based projection is suggested to better model the encoded utterances.

Kim et al. [141] propose a two-step ANN-based dialog state tracker, which is composed of an informativeness classifier and a neural tracker. The informative CNN-based classifier filters out non-informative utterances, and the neural tracker estimates dialog states from the remaining informative utterances.

Mrksic et al. [197] consider the issue of developing a state tracker for goal oriented CAs. They consider the difficulty of scaling the state tracker to large and complex dialogue domains because of the dependency on large training sets. They propose a neural belief tracking (NBT) framework that uses pretrained word embeddings to learn the distribution of user contexts.

Su et al. [270] estimate the task success by inspecting the dialogue as it evolves, by utilizing RNNs and CNNs. Their experiments demonstrate that both RNNs and CNNs can accurately estimate when substantial training data are available, though RNNs are more robust when training data is limited. Many goal oriented CAs are trained on available goal oriented datasets, (see Section 3.7 for more details on such datasets). Other goal oriented CAs are trained on human users. While such training may yield richer dialogues, it is more expensive.

Liu and Lane [169] address the challenges of building a reliable user simulator to train a goal-oriented CA by simulating the dialogues between two agents. Initially, a basic conversational agent and a basic user simulator are trained on dialog corpora through supervised learning, and then their abilities are improved by allowing them to conduct task-oriented dialogues while iteratively improving the policies using DRL.

3.4 Human Related Issues

In addition to the technical issues of natural language understanding and generation, good conversational agents should be aware of human characteristics, observe user emotions, provide empathy in their responses, and engage the user.

According to Clark et al. [68], humans perceive the communication with CA as a means to achieve functional goals. In their study, Clark et al. present the results of semi-structured interviews on how people view the conversation between humans and CAs. They find that several social features reported as crucial in human-human conversation, such as understanding and common ground, trust, active listenership, and humor, are not listed as required for human-CA conversations. CA conversations are described almost exclusively by transactional and utilitarian terms. However, this view of CAs is not satisfactory in domains that require the user to engage and form an emotional bond with the CA.

Yand et al. [305] argue that understanding users' affective experience is crucial to the design of compelling CAs. To elaborate on this claim, they survey 171 CA users of Google assistant, and examine the affective responses in four major usage scenarios. In addition, they observe the factors that influence affective responses. They find that the overall experience of the user is positive, with the most salient emotion being interest.

Both pragmatic and hedonic qualities influence affective experience. The factors underlying the pragmatic quality are helpfulness, proactivity, fluidity, seamlessness, and responsiveness. The factors underlying the hedonic quality are comfort in human-machine conversation, pride of using cutting-edge technology, fun during use, perception of having a human-like assistant, concern about privacy, and fear of causing distraction. In the remainder of this section, several issues are discussed that can assist in establishing a deeper connection between the user and the CA during conversations. The focus is on the following aspects: emotional issues, CA personality, and adaptation to the taste and needs of the user.

Emotional Aspect of Conversations

Emotional understanding and empathy are important abilities for CAs acting in several social domains including health care, education, and customer support; however, these abilities are also useful to CAs, in general. Combining emotional awareness with technologies and methods for CAs, requires multi-domain knowledge in psychology, artificial intelligence, sociology, and education research.

The challenge in enabling empathy and emotionally adjusted responses is twofold: first, the agent must be able to detect the emotional state of the human; second, it must be able to provide the proper emotional response.

The agent may be able to detect user emotions based on user utterances as well as voice and body language. Emotion detection (ED) is an important branch of sentiment analysis and deals with the extraction and analysis of emotions from text and from audio. A detailed review on emotion recognition technologies and studies is provided in Section 2.2.

Emotional effects, as well as properties of the speaking style, can be added to the CA to generate speech which is closer to human dialog. The ability to recognize the emotions and feelings of others, and replying accordingly is known as empathy, which is a crucial socio-emotional behavior for smooth interpersonal interactions. Empathy can be verbal and non-verbal. Yalcin [302] suggests that embodied CAs should be equipped with real time multimodal empathic interaction capabilities. The empathic framework leverages three hierarchical levels of capabilities to model empathy for CAs. Following the theoretical background on empathic behavior in humans, the embodied CA can express empathy by using facial expressions,

gaze, head, and body gestures as well as verbal responses.

Tellols et al. [226] propose equipping the CA with sentient capacities, using ML technologies. They illustrate their proposal by embedding a virtual tutor in an educational application for children. Their CA has a unique personality, emotional understanding, and needs that the user has to meet. The CA's needs can be expressed by Maslow's hierarchy of needs [185]. Tellols et al. tested the two CA versions with 10–12 year-old students and found that the second version, equipped with ML capabilities, displays higher understanding capacity and yields a nearly 100% user satisfaction rate.

To summarize, considering that the user's emotional experience and engagement are of great importance in various social and health domains, several studies suggest methods to recognize user's emotional state to provide an appropriate empathic response. The emotional awareness of CAs can make the user more satisfied and can yield longer and meaningful human-CA conversations.

The Effect of CA Personality

Recent studies have observed that adding personality aspects and human-like characteristics to the conversation may strengthen the connection of the user with the CA. In particular, in the mental health care domain, such CAs can elicit higher engagement from humans during the therapeutic process.

Chavesa and Gerosa [64] survey 56 papers from various domains to understand how social characteristics in CAs benefit human-CA interactions. They define eleven social characteristics: proactivity, conscientiousness, communicability, damage control, thoroughness, manners, moral agency, emotional intelligence, personalization, identity, and personality, further grouping them into three social categories: conversational intelligence, social intelligence, and personification. They show that certain characteristics, such as moral agency and communicability are influenced by the domain, while others, such as manners and damage control, are more generally applicable. They further point out that social science theories, such as cooperative principle and mind perception theories, can contribute to the design of CAs with social characteristics.

Zhang et al. [315] propose endowing CAs with a profile of configurable, yet persistent, persona to make them more engaging. This profile is encoded by multiple sentences of textual description. To train the CAs on personal topics, they present a new dialogue dataset consisting of 164,356 utterances between crowd workers who were asked to chat naturally to get to know each other during the conversation.

Inspired by the vision of human-like interactions of conversational agents, Volkel et al.

[289] examine the important features of CA's personality. They use various sources to examine the main adjectives used by CAs, including an online survey, an interaction task in the lab, and a text analysis of 30,000 online reviews of CAs. They aggregate the results into a set of 349 adjectives, which were rated by 744 people in an online survey. A factor analysis reveals that the commonly used big five model for human personality [243] does not adequately describe the CA personality. As an initial step in developing a personality model, Vokel et al. propose an alternative set of main features to be applied to the design of CA personalities.

Feine et al. [91] observe the process of how a social cue evolves into a social signal and subsequently triggers a social reaction. Using the theory of interpersonal communication [55], they identify a taxonomy of social cues of ECAs and classify the social cues into four major categories and ten sub-categories. The four major categories are: verbal, visual, auditory, and invisible. They evaluate the mapping between the identified social cues and the categories, using a card sorting approach.

The effect of ECA personas and cues on user engagement was studied by Liao and He [164]. In their experiment, participants were randomly assigned to racial-mirroring ECAs, non-mirroring ECAs, or control groups. After interacting with the ECA, participants completed a survey assessing their perception and evaluation of the agent. Liao and He demonstrated that racial mirroring has a positive influence on the user's perceived interpersonal closeness with the agent, and the participants interacting with mirroring ECAs reported a higher level of satisfaction and a higher desire to continue interacting with the agent, and predicted a closer future relationship. In addition, people were significantly more likely to select same-race agent personas when they were given an opportunity to customize the ECA.

Go and Sundar [103] tested the distinct and combined effects of three types of cues that potentially enhance the humanness of chat agents: human-like visual cues, the use of human names or identities, and the use of human language. For these three factors, the authors examine how interactions among these cues influence psychological, attitudinal, and behavioral outcomes. Their experimental results indicate that CA interactivity is an important factor in determining psychological, attitudinal, and behavioral outcomes, while the identity cue turns out to be a key factor in eliciting certain expectations regarding CA's performance in conversation. However, message interactivity can compensate for the impersonal CA nature.

A good open-domain CA should be able to seamlessly blend all its skills, including the ability to be engaging, knowledgeable, and empathetic into one conversational flow. Smith et al. [263] present a method for training a CA with blended skills and testing it. They show that existing single-skill tasks can effectively be combined to obtain a model that blends all skills into a single CA. To preclude unwanted biases when selecting the skill, fine-tuning is done on the blended data.

Personalized CAs and their Effect on Human Engagements

In addition to possessing empathy, persona, and knowledge, the ability of the CA to adapt itself to the user's taste and needs, is also important in engaging the user.

The studies described in this section, are related to personalized CAs that adapt themselves to particular users to increase user satisfaction. However, adaptation may come at the cost of a loss in user privacy, which, if observed by the user, may limit the user's spontaneity in conversation. The effect of users limiting their conversation, upon detecting that the CA is collecting private information to adapt, was reported by [93].

A psycholinguistic characteristic of young adults interacting with a CA, is to discuss daily scheduling concerns and stress levels. Ferland and Koutstaal performed a linguistic analysis that presents the slightly paradoxical effect of reduced user engagement when a conversational agent explicitly discloses information on its user model to the user. They conclude that overt user models may discourage users from self-disclosure and participation in an information-rich spontaneous conversation.

Nevertheless, in task-oriented domains as well as educational domains, adaptation to the user's abilities and skills may assist the CA to be more effective and may result in higher user satisfaction. Carfora et al. [62] envisage goal-oriented agents whose policies take into consideration the psychological features of the user to deliver personalized and more effective messages. They built a probabilistic predictor based on the theory of planned behavior [15] and psycho-social model of reference and implement it by a dynamic Bayesian network.

The smart learning environment may involve task assignments adapted to the learner's abilities [241], smart hints and feedbacks [34], smart guidance during the learning process [281], and personalized conversational agents who assist in the learning process [296].

In the healthcare domain, Mandy [202], a primary care CA created to assist healthcare staff by automating the patient intake process, provides personalized intake service to patients by understanding their symptom descriptions and generating corresponding questions during the intake interview.

Schuetzler et al. [255] focus on the effect of improving the social presence of CAs by enhancing their responsiveness and embodiment. Responsiveness is the ability of the agent to provide responses contingent on user messages, and embodiment is the visual representation of the agent. In particular, they examine the influence of CA responsiveness and embodiment on the answers people give in response to sensitive and non-sensitive questions. They find that CA responsiveness increases socially desirable responses to sensitive questions.

Figure 8 presents an overview of the human related issues discussed in this section. Each

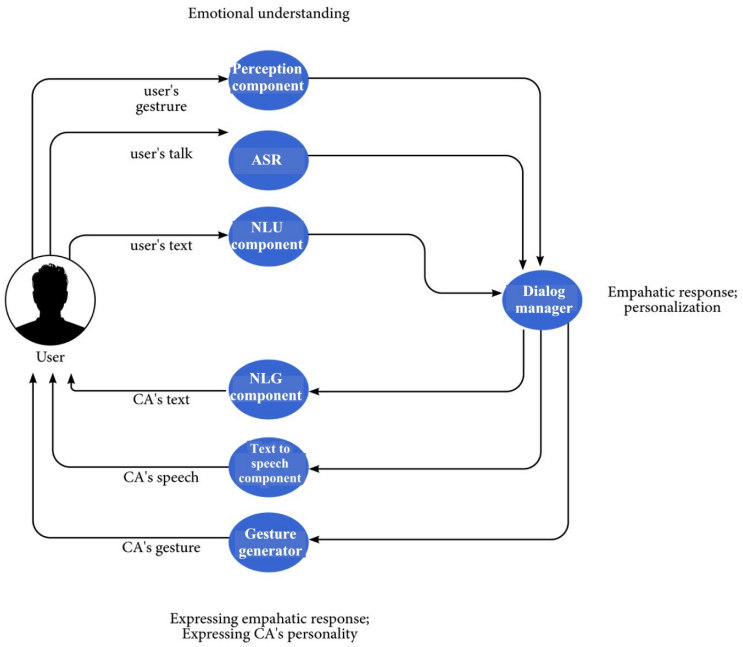

Figure 8: Human related aspects of the CA: emotion sensitivity, personality expression, and adaptation to the user's taste and needs.

challenge is associated with the appropriate CA component expected to assume the most responsibility for that challenge. Understanding the user emotional state is mostly a challenge of the ASR, NLU, and perception components; the dialog manager decides on how to provide an appropriate empathic response; the NLG, the gesture generator, and the text-to-speech components are responsible for generating empathy in verbal and non-verbal responses; the personality of the CA is expressed by the response generators including the text generator, speech generator, and gesture generator components; and adaptation of the CA to the user's taste and needs is the responsibility of the dialog manager.

3.5 Goals and Applications of Conversational Agents

Personal Assistants and Open Domain Conversational Agents

The first CA was developed in 1964 by Weizenbaum [293]. It was named ELIZA, and it simulated conversations by using a pattern matching approach. ELIZA was designed to serve as a psychologist and mimicked certain kinds of natural language conversation between

humans and computers. People mistakenly believed ELIZA to be intelligent enough to comprehend a conversation, and some even became emotionally close to it. In 1972 the psychiatrist Kenneth Colby developed PARRY [69], which is a natural language program that simulates the thinking of a paranoid individual. PARRY was developed to train users to detect people at psychological risk.

DeepProbe [307], RubyStar [171], and Meena [9] are recently developed open domain chatbots. DeepProbe uses a sequence-to-sequence mechanism to satisfy user queries. RubyStar combines ML models and template and rule-based responses; it uses topic detection, engagement monitoring, and context tracking. Meena CA is trained end-to-end on data mined and filtered from conversations on social media.

Currently, mobile devices and smart speakers are equipped with powerful agents such as Siri, Cortana, Alexa, and Google Assistant, offering support for a variety of tasks such as question answering, information retrieval, scheduling meetings, sending messages, and controlling smart home devices [80, 126]. These assistants constantly listen to hear a wake-up keyword, for example, "Okay Google", "Alexa", etc. Once a wake-up keyword is said, the assistant records the user's command and sends it to a server. The server translates the voice command to text by using an ASR component that parses the text using a parser and uses a natural language understanding component to determine the appropriate response or action to be taken by the assistant. For example, a simple query "How are you today?" may be followed by an answer "I'm fine; thank you.". A more sophisticated question, such as "How many types of mammals are there?", may invoke a web-search which results in an answer such as "There are 6,000 different species of mammals". Commands requesting turning on the lights, setting the temperature of an air conditioner, playing a specific song, or ordering a product are executed accordingly.

Current virtual assistants have several drawbacks. First, they require a steady internet connection. Second, while they usually support multiple languages, they are far from supporting all languages used world-wide. In addition, virtual assistants that order products or book hotels and flights may cause unintentional expenses e.g., when the user is a child. Misinterpretation may cause the virtual assistant to send an unwanted message. This may be harmful if the wrong message is sent to the wrong person, or if a conversation is unintentionally recorded and sent to the wrong person. A virtual assistant may also enable the installation of malware. Misinterpretations may also cause the accidental turning off of the heating in a house with a baby, which may have devastating consequences.

Some virtual assistants give programmers the ability to extend their abilities. For example, Alexa allows programmers to extend her abilities using the Alexa Skill Kit (ASK). Participants in the Alexa prize challenge developed social chatting skills for Alexa. There are few open

domain CAs that enable a lay user, rather than a programmer, to teach the agent to perform new action sequences or new responses. Learning by instruction agent (LIA) [31] uses a combinatory categorial grammar (CCG) semantic parser to transform the semantics of each command to a few terms of primitive executable procedures which define sensors and effectors of the agent. If the user gives LIA a natural language command and if the LIA does not know how to execute the command, it will ask the user to explain how to realize the command through a sequence of natural language steps. Once explained, the LIA can execute the command in the future.

SUGILITE [161] is a programming by demonstration (PBD) system that uses the Android's accessibility API to enable users to create automation on smartphones. In case the user specifies commands that SUGILITE does not know how to execute, it prompts the user to demonstrate the command, records the user's explanation, and automatically generates a script. Thus, SUGLITE can learn to execute an unrecognized command from a single demonstration.

Safebot is a collaborative chatbot that allows users to teach the agent new responses [189]. Safebot allows the users to identify inappropriate responses, which are then removed from Safebot's database such that future users are not allowed to teach Safebot responses similar to the ones previously tagged as inappropriate.

KBot [14] is a comprehensive open-access CA that exploits the potential of semantic web technologies, federated databases, and NLU. KBot contributes to a better understanding of user queries in the context of linked data by being able to answer different user queries. It can handle tasks such as conversations in English, social network conversations, FAQs, and mathematical tasks, using information gathered from multiple sources such as DBpedia, Wikidata, and MyPersonality [3] datasets.

Finally, MILABOT [258] is a DRL-based CA, developed for the Amazon Alexa Prize competition. MILABOT is capable of chatting with humans through speech or text. It was trained on crowdsource data and real-world user interactions.

Educational Applications

Online learning has shown significant growth over recent years, in particular, during the COVID-19 outbreak. Unfortunately, in online learning, teachers and students are distant from each other, and therefore, the connection and interaction between them may be insufficient. This may cause online learning to be less effective.

There have been multiple attempts to enhance online learning by using intelligent tutoring

[3] `http://mypersonality.org`

systems (ITS) [212], which are customized, computer-based instruction and feedback methods without human intervention. Many include conversational agents, which can interact with the students in natural language during the learning process.

Paschoal et al. [219] survey 101 pedagogical conversational agents. They identify the different educational areas for which conversational agents have been developed, discuss common development techniques for pedagogical CAs, and also survey the communication strategies used by pedagogical CAs to interact with students, Some successful CAs that are recently used in the education domain are next described. Sara is a CA to assist students with learning [231]. Sara shows online video lectures and asks questions to ensure that the student has understood the lecture. It offers additional information and explanations if the student's responses are inaccurate. Sara interacts by voice and text when needed and has a voice-based input mode. It was demonstrated to improve learning in a programming task. A similar CA was developed by Paschoal et al. [220] to support software testing. AutoTutor [107] is a computer tutor that simulates the dialogues and strategies of a human tutor. It presents questions and problems from a curriculum script, and according to the learner's input, decides which action to perform next (e.g., providing a hint, moving on to the next problem). AutoTutor segments the input from the learner into a sequence of words, to assign alternative syntactic tags to words and the correct syntactic class to a word.

MSRBot is a question answering CA, dedicated to software related issues [3]. It uses a neural network to classify each speech act into one of five speech act categories: assertion, wh-question, yes/no question, directive, and response. It extracts useful information from software repositories to answer several common software development/maintenance questions.

Hobert [124] presents the design and evaluation of a chatbot-based tutor to help teach beginner programmers to code in university courses. Hobert's coding tutor is based on teaching assistant requirements that appear in the scientific literature. Hobert claims that his chatbot tutor is suited to take over the tasks of teaching assistants when there is no human teaching assistant available.

Similarly, Kloos et al. and Aguirre et al. [145, 12] introduce the design and features of a CA for Google Assistant [28] to complement a massive open online course (MOOC) for learning Java. Both studies run several experiments and report that users find the conversational agents to be very useful.

Lin et al. [167] developed Zhorai, a CA that enables children to explore AI algorithms and machine learning. Lin et al. show that by training an agent, observing its mistakes, and retraining the agent, children were able to understand the agent's ability to learn, as well as obtain some level of understanding of the learning algorithms used by it.

Cai et al. [60] introduced MathBot, a rule-based chatbot that explains math concepts,

provides practice questions, solves problems, and offers tailored feedback. Using mTurk workers, Mathbot was compared to other baseline methods, such as video tutorials and written material. It was found that students prefer MathBot over other options.

CAs can also be useful in foreign language learning. Indeed, there have been several recent attempts to develop CAs for that purpose. Duolingo's chatbot with Mondly as well as Andy are some examples of chatbot applications for language learning [143]. Some virtual assistants, such as Alexa, include extensions that enable the learning of foreign languages [179]. Alexa has the skills to assist in building a vocabulary and handling a conversation in a foreign language. Pham et al. [224] develop English Practice, which is a mobile chatbot application to assist a user in learning new vocabulary and to carry on a conversation. Another CA dedicated to language learning is Lucy [90], an embodied virtual agent, designed to help users to learn vocabulary and grammar and to carry on a conversation.

CAs can also be used to support the administration in educational systems. For example, Hien et al. [121] present FIT-EBot, a chatbot that responds to student questions related to services provided by the education system on behalf of the academic staff. Similarly, Ranoliya et al. [235] introduce a chatbot designed to answer visitor questions at Manipal University. It provides an answer based on a dataset of frequently asked questions (FAQ) using AIML. When a user asks a query, the chatbot searches for a similar question and provides the answer to that question. Another chatbot was developed by Keeheon et al. [152] to provide information in educational systems by answering frequently asked questions The chatbot was successfully used by students and department offices in Underwood International College, Korea.

The authors reported that the use of the chatbot had a positive influence on administrative work in reducing workload.

Discussion-bot [92], developed by Feng et al., provides answers to students' discussion board questions using natural language. Given a question, it mines suitable answers from an annotated corpus of archived discussions and course documents and chooses an appropriate response. For a review of conversational agents and other tools which assist children and adults with special needs see 2.1

Healthcare Conversational Agents

CAs can potentially play an important role in health care. There have been several recent reviews on CAs in this field (see [151, 61, 276, 195]). Each points to challenges in the healthcare area pertaining to efficiency, security, and privacy.

CoachAI is a system that includes a chatbot and a machine-learning model to support a patient's health activities [89]. The chatbot collects data, sends reminders, and converses with

users through text-based, simple, graphical elements to guide the user in health related issues. The model is based is real-world data provided by a health clinic. The application provides the caregivers with insights on the users, and assists with the tracking of user activities and their health conditions.

Daily healthcare can be overwhelming for people with a chronic disease. Neerincx et al. [200] developed a social robot that helps children with diabetes. The robot supports the daily diabetes management processes, namely, taking pills, shots, and body measurements by conversing with the child.

Watson assistant for health (Watson Health) is an extension of IBM Watson [122] to the healthcare domain. Watson was originally developed for the Jeopardy challenge. Watson Health [269] is a CA for health support. It uses a text-based natural language interface. It receives a collection of patient symptoms and produces a list of possible diagnoses. The assistant provides detailed annotation as well as links to supporting medical literature. However, a study conducted by Ross and Swetlitz [245] indicates that in some cancer cases, Watson Health provided unsafe and incorrect recommendations.

Xu et al. [301] introduced KR-DS, a chatbot for the healthcare domain. KR-DS obtains a set of symptoms from the user, recognizes the bio tags of each word using Bi-LSTM, classifies the intent of each sentence, and finally, provides a diagnosis to the user, in natural language, using a medical knowledge graph. Experiments show that KR-DS outperforms other state-of-the-art methods in diagnosis accuracy.

Fitzpatrick et al. [96] developed Woebot, a medical voice-based CA for cognitive-behavioral therapy dedicated to nonclinical cases addressing low mood and anxiety. Woebot provides mental health information, recommends activities for specific mood problems, and handles emergency support services. The users reported an improvement in their mood after using Woebot.

Edwards et al. [85] introduced Tanya, a graphically embodied female agent that supports breastfeeding. Tanya was deployed in a hospital and was accessible to women after birth. Edwards et al. show that women that interacted with Tanya increased their chance of successful breastfeeding for the first six month.

During the COVID-19 outbreak, people require medical information with respect to the outbreak but cannot obtain the information from medical teams, which are overwhelmed. Yang et al. [304] developed a medical chatbot that can be consulted for COVID19-related issues. The chatbot is trained on two datasets, in English and Chinese, containing conversations between doctors and patients on COVID-19.

Despite all the CAs developed in the field of healthcare, the reception of CAs in this field has not been as positive as expected. Palanica et al. [213] examine the perspectives of

practicing medical physicians on the use of healthcare CAs for patients. Their results indicate that many physicians believe that CAs would be most beneficial for scheduling doctor appointments, locating health clinics, and providing medication information. However, most of the physicians believe that CAs cannot effectively take care of patients' needs or provide detailed diagnosis and treatment. Nadarzynski et al. [198] study the acceptability of CAs in healthcare from the perspective of the general public. While the participants in the study recognize the potential of CAs in healthcare, they state that their experience is not satisfactory enough, and that they are concerned about security issues. Scholten et al. [253] survey several CAs in the field of healthcare. They conclude that while CAs can increase the motivation of patients and promote behavioral change, user needs are many times implicit, and these needs cannot be addressed by CAs.

CAs in the Business Domain

Conversational agents are becoming more and more prominent in a diverse range of applications in the business area. According to Dhanda [79], CAs have reduced costs in organizations by approximately $48.3 million in 2018 and are expected to reduce costs by $11.5 billion by 2023. See Bavarescoa et al. [41] for a literature review on CAs in the business domain with a focus on machine learning. CAs can be used as customer service assistants, providing answers to frequently asked questions (FAQs), which is a common task that can be handled by CAs.

The Thomas question answering chatbot [277] uses artificial intelligence markup language (AIML) for template-based questions like greetings and general questions and latent semantic analysis (LSA) [277] to answer other related questions. If the chatbot cannot find a relevant answer, it asks the user for a clarification.

Another chatbot in the customer service area is SuperAgent [71], which leverages large-scale and publicly available ecommerce data. Given a user request for information about a specific product, SuperAgent provides relevant information from in-page product descriptions and from ecommerce websites. SuperAgent is provided as an add-on extension to Microsoft Edge and Google Chrome browsers.

Xu et al. [300] created a chatbot to serve users' requests on social media (Twitter). The chatbot encourages interaction between users and businesses on social media. The chatbot was trained on nearly one million Twitter conversations between users and agents. Their analysis indicates that over 40% of user requests are emotional and do not intend to seek specific information. They show that their chatbot, which is based on deep learning, yields a higher BLEU score [215] than that of an information retrieval-based system.

Yan et al. [303] introduce a chatbot, dedicated to online shopping. The goal is to assist online customers in purchase-related tasks by answering specific questions and searching for a product. They integrate this system into a mobile online shopping application with millions of consumers.

Another chatbot is SamBot [225], which is integrated into Samsung's website to answer user questions. Its knowledge base includes: Samsung promotion, Samsung product FAQs, and general information related to Samsung (e.g., open hours and branch locations). If a proper answer cannot be found, SamBot generates a random answer. It can also recommend users questions to ask. They show that SamBot is capable of handling Samsung related questions very well.

Kaghyan et al. [138] review the aspects of business-to-business (B2B) tools including the use of CAs. In their article, they describe several ways and platforms for creating Facebook chatbots that support a business. Detailed descriptions are provided for three chatbot creation platforms: Chatfuel, ManyChat, and "It's Alive!", and a comparison is performed with respect to capabilities, strengths, and limitations.

Another use of CAs in the business domain is for negotiation. Lewis et al. [155] demonstrate that it is possible to train end-to-end CAs for negotiation, which is simultaneously a linguistic and a reasoning problem. To achieve this goal, their CAs contain adversarial elements as well as cooperative elements, and the CAs are required to understand, plan, and generate utterances. They collected a dataset of natural language negotiations between two people to show that their end-to-end neural models successfully imitate human behavior in this domain.

Luo et al. [176] collaborated with a large financial service company, to design a randomized field experiment on the consequences of chatbots hiding or revealing that they are indeed chatbots. They conclude that when the true identity of chatbots is not disclosed, CAs are as effective as proficient workers and four times more effective than inexperienced workers in increasing customer purchases. However, when chatbots disclose their identity before conversation, the purchase rates are reduced by more than 79.7% and the conversation becomes shorter. Unfortunately, users do not always trust that CAs can provide the required support.

Følstad et al. [97] present an interview study of thirteen users who interact with chatbots in customer support regarding their experience and the factors affecting their trust. The users' trust was found to be affected by different attributes such as the quality of the CA's interpretation of the requests and whether the generated text seemed human-like.

Chihsun et al. [156] investigated how users cope with conversations with chatbots that do not make any 'progress, in the field of customer support. They analyzed a three-month conversation log with a chatbot, which was taken by one of the top digital-banking institu-

tions in Taiwan. They found 12 types of conversational non-progress and 10 types of coping strategies on the part of the user.

Abdellatif et al. used Google's Dialogflow engine [105] to extract the user intent and the entities mentioned in the user input, Their initial training set was collected from a group of software developers and consisted of different ways developers pose similar questions. Additional training data were collected from developers using the initial CA version during a test period.

Influence and Malicious CAs in Social Networks

Several conversational agents are developed for deployment in social networks. These CAs attempt to influence public opinion by persuading specific surfers to take certain actions, consume certain products, or influence political views.

Few internet tutorials [10, 74] have been written to guide users in the process of Twitter chatbot development. Adams [8] gives an overview of influence impersonating CAs, which impersonate a human to influence users on social media. They also state that most impersonator chatbots are very simple and therefore, cannot deceive serious interrogators.

The study of Assenmacher et al. [27] provides insights into markets of influence and malicious chatbots as well as an analysis of freely available software tools, which are used to create them. Similar to Adams, they conclude that current influence chatbots are very simple and, despite the major advances in the literature on CAs, still use very simple automation methods.

Another study in the social chatbot area is that of Kollany [147]. According to Kollany, there is an exponential growth in the number of influence chatbots on Twitter. Kollany gathered data from GitHub on the ways developers collaborate with each other and check social aspects of programming on that platform.

While influence CAs are usually intended only to influence a person's opinion, some malicious CAs utilize a social network to steal personal and private information including credit card and bank account details, or to spread false information in an attempt to manipulate the stock market [95].

Several studies focus on influence and malicious chatbots acting in social media. Varol et al. [285] use a publicly available dataset of Twitter accounts and manually label all users either as humans or influence chatbots. They estimate that 9-15% of active Twitter accounts exhibit influence chatbot behavior. They present a machine learning model to detect influence chatbots on Twitter based on features extracted from the dataset, such as user followers and tweet content and sentiment.

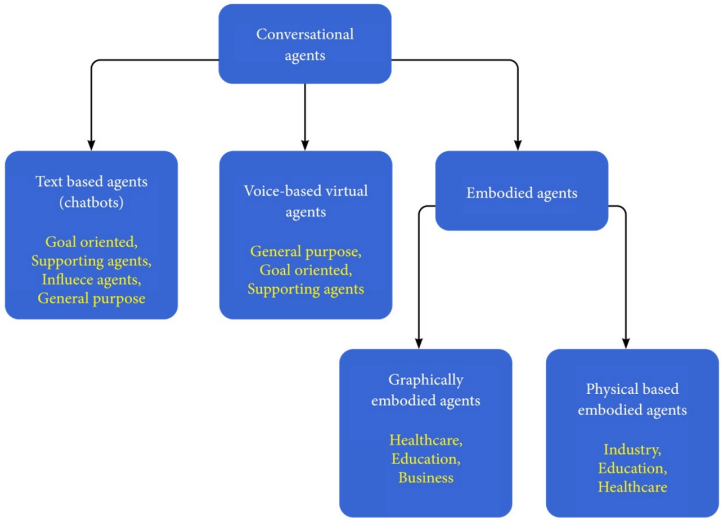

Figure 9: Conversational agent applications

DARPA held a 4-week competition in 2015 in which multiple teams competed to detect influence chatbots on Twitter [271]. Out of 7038 Twitter accounts, 39 were labeled by DARPA as influence chatbots. The leading group detected all influence chatbots, using a combination of machine learning techniques along with a user support system.

Lee et al. [153] deployed honeypots in the Twitter social network to identify and analyze content polluters. They investigate the attributes of Twitter users, including user behavior over time, user followers, and user following. They also enumerate features that may assist in identifying content polluters automatically, and present a classification model. Finally, they show that their model successfully identifies content polluters.

To summarize this section, Fig. 9 refers to the CA definitions (provided in Fig. 1), and for each type of CA, details the domain of applicability.

3.6 Evaluation Metrics

Three main approaches are used in the literature for evaluating the quality of a conversation agent: human based evaluation procedures, machine evaluation metrics based on language characteristics, and an ML approach trained on a dataset consisting of human evaluations. The advantages of human evaluation are clear, as humans can evaluate whether the CA responses seem appropriate and resemble responses. However, since human evaluation proce-

dures are expensive, several automatic methods have been proposed for the evaluation process. In addition, it should be noted that the judgment of the human evaluators is subjective, and can be affected by external conditions, such as mood, and this makes it difficult to rely on one person's judgment. A team of examiners is an effective approach to handle this difficulty but, clearly, this solution increases the evaluation expenses accordingly. Another possibility is to use automated tools for evaluation. Unfortunately, due to the linguistic richness of natural languages and the wide variety of reasonable response options, it is still challenging to achieve accurate and meaningful evaluation when using automatic tools. Therefore, the ML approach tries to benefit from both approaches; on the one side it is based on human evaluation, and on the other side, it does not require new implicit costly evaluation methods for each new dialog situation.

Radziwill and Benton [230] present a literature review of quality issues related to CA development and implementation, focusing on two topics: quality attributes and quality assessment approaches. Deriu et al. [77] survey the main concepts and methods of CA evaluation. For each type of CA, task-oriented, conversational, and question-answering dialog systems, they define the main technologies and the evaluation methods that are appropriate for that type. The requirements of the evaluation methods are stated with respect to automated or partially automated evaluation, repeatability of the results, correlation with human judgement, ability to focus on CA features, and explainability. Finally, Masche and Le [183] divide the different evaluation methods into four classes: qualitative analysis, quantitative analysis, pre/post-test, and CA competition.

In this section, the evaluation methods are divided into three classes, according to the way they are obtained, namely human-based evaluation, machine-based evaluation, and the ML approach, and some popular evaluation methods are further described for each of these three classes.

Human Based Evaluation Procedures

As mentioned above, the most accurate method to assess the dialog quality of a CA, is through the score and the qualitative description obtained from humans interacting with the CA. Deriu et al. [77] describe various approaches of human evaluation consisting of lab experiments with users invited to interact with a CA and subsequently asked to fill out a questionnaire; in-field experiments with feedback collected from real users of the CA; and crowdsourcing with crowd workers, either asked to talk to the CA and then rate it or asked to read a produced dialogue and then rate it. The CA rating is based on quality, fluency, appropriateness, and sensibleness.

Venkatesh et al. [286] describe the following metrics to evaluate an open-domain CA: user experience, coherence, engagement, domain coverage, topical depth, and topical diversity. In addition, they propose a unified evaluation strategy, which combines the above metrics into a new evaluation model that correlates well with human judgement. Their unified evaluation strategy was applied throughout the Alexa Prize competition to select the top performing CAs.

Griol et al. [109] define a set of specific measures to evaluate the quality of a medically oriented CA. The proposed measures are divided into high-level dialog features, dialog style, and cooperativeness. High-level dialog features evaluate how long the dialog lasts, how much information is transmitted in individual turns, and how active the dialog participants are, while dialog style and cooperativeness features analyze the contents of different speech actions.

To summarize, there are generally three main sources of human based evaluation: lab sources, real CA users, and crowdsourcing. The information obtained from humans can include: qualitative and quantitative questionnaires, real CA user feedbacks, and dialog features.

Machine Evaluation Metrics

Since a high cost is associated with human evaluation, machine-based evaluation or hybrid human-machine-based evaluation are widely used to examine the quality of CAs. Machine-based CA evaluation is challenging due to the lack of an explicit objective for conversation performance measurement. Several studies utilize machine translation-based metrics for CA quality evaluation.

One such metric is the BLEU score [216], a text summarization metric developed for automatic evaluation of machine translation. BLEU takes the geometric mean of the test corpus modified precision scores and multiplies it by an exponential brevity penalty factor. The main component of BLEU is the n-gram precision, which is the proportion of the matched n-grams out of the total number of n-grams in the evaluated translation.

Recall-oriented understudy for gisting evaluation (ROUGE) [166], originally developed for automatic summarization, is also adapted to CA evaluation. Similar to BLEU, ROUGE counts the number of language units, such as n-grams, that appear both in the evaluated summary and in the ideal human-generated summary.

Another popular evaluation metric for machine translation that is applied to CA evaluation is METEOR [38]. METEOR evaluates a translation by counting word-to-word matches between a translation and the reference sentence. If more than one reference is available, the given translation is scored against each reference independently, and the best score is

reported.

Liu et al. [170] investigate the usage of the above translation and summarization evaluation metrics for CA. They note that available machine translation metrics assume that valid responses should have significant word overlap with the ground truth responses. This is a strong assumption for CAs, which exhibit a significant diversity in the space of valid responses. They show that many commonly used metrics for CA evaluation do not correlate strongly with human judgement, and they conclude that there is a need for a new metric that correlates more strongly with human judgement.

Machine Learning Based Evaluation

A third approach of CA evaluation is to use ML to predict the human rating of CAs' dialogues. Lowe et al. [174] present a dialogue evaluation model called ADEM that learns to predict human-like scores for CA responses, using a dataset of human scores of responses. The human scores were collected using crowd workers that were shown a dialogue context and a candidate response and asked to rate the responses. ADEM is trained by an RNN and, given a response, can successfully predict the appropriateness rating of the response as if it is a human.

Tao et al. [272] propose a routine for evaluating system responses called RUBER. RUBER consists of a Siamese neural network, trained to predict if a pair of context and response are relevant. RUBER is trained using two metrics: a referenced metric measures the similarity between the generated response and the ground truth response, and an unreferenced metric measures the relatedness between the generated response and the original query. The referenced and unreferenced metrics are combined with heuristic strategies (e.g., averaging) to further improve RUBER's performance.

Guo et al. [113] propose a topic-based evaluation method on topic breadth, which checks the ability of the CA to talk about a large variety of topics, and topic depth, which checks the ability of the CA to handle a long and cohesive conversation about one topic. A deep average network (DAN) was used to train the topic classifier on a variety of questions and query data, categorized into multiple topics. To summarize, the ML approach of evaluation can be helpful to a wide range of CA researchers and developers as it combines the advantage of human judgement with the advantage of resource saving to rate an unlimited number of CAs and dialogues, utilizing the trained evaluation model.

Tables 2 and 3 provide the technologies and the evaluation method(s) behind each of the main CAs described in Section 3.5.

Table 2: Technologies and evaluation methods for main CA applications: part A

Personal Assistants and Open Domain CAs			
CA	**Short description**	**Main technology**	**Evaluation Method**
ALICE [290]	a general purpose chatbot	AIML, pattern matching	winner of "the most human computer" 2000,2001,2004
LSA-bot [11]	ad-hoc implementation of the LSA framework	Latent Semantic Analysis (LSA)	-
IRIS [37]	example based chatbot	vector space model cosine similarity metric	success and failure examples
DeepProbe [307]	an open domain chatbot chatbot	seq-2-seq	AUC scores
RubyStar [171]	an open domain chatbot	seq-2-seq, topic detection, engagement monitoring, context tracking	human evaluation by the Alexa Prize evaluation
Siri [50]	Apple's virtual assistant	CNN, LSTM	commercial application
Cortana [45]	voice controlled assistant for Microsoft windows	NLP, Tellme Networks, Semantic search database	commercial application
Alexa [173]	Amazon voice assistant	NLP, LSTM	commercial application
KBot [14]	knowledge chatbot	SVM+analytical queries engine	F-score, precision, recall, intent classification
MILABOT [258]	speech/text CA	DRL	Amazon Alexa Prize competition

Discussion-Bot [92]	question answering chatbot	semantically-related matching, TF-IDF metric	human judges classified the answers quality

Goal Oriented CAs			
CA	**Short description**	**Main technology**	**Evaluation Method**
SUGILITE [161]	Programming-By-Demonstration system	frame-based dialog management	a lab study: task completion time
Safebot [189]	collaborative chatbot	parser+Word2Vec	users' engagement
LIA [33]	learning by instructions agent	uses combinatory categorial grammar (CCG) parser	speed of task completeness

CAs for Social Support			
CA	**Short description**	**Main technology**	**Evaluation Method**
ELIZA [293]	the first CA: emulates a psychologist	pattern matching	people experience
XiaoIce [321]	a popular social CA	IQ + EQ + Personality	human rating
Meena [9]	a sensible chatbot	generative chatbot trained end-to-end on social media conversations	human evaluation metric called Sensibleness and Specificity Average (SSA)

Table 3: Technologies and evaluation methods for main CA applications: part B

Educational CAs			
CA	**Short description**	**Main technology**	**Evaluation Method**
Sara [231]	student's assistant	scaffolding strategy	pretest and posttest scores of learners pro-survey and post-survey
AutoTutor [107]	computer tutor	LSA, pattern maching speech act classification	learning gain
MSRbot [3]	sofware related Q&A	Dialogflow	effectiveness, efficience
Zhorai [167]	CA for children to explore ML concepts	NLTK package Website visualizer	accuracy, child's level of engagement
MathBot [60]	math teaching chatbot	rule based	crowd worker preferences
English Practice [224]	Personal Assistant for Mobile Language Learning	Dialogflow platform	statistics about real users
Lucy [90]	embodied on-line virtual agent for language learning	ALICE offshoot	demonstrative examples
FIT-EBot [121]	administrative chatbot	DialogFlow	students reports
QTrobot [86]	social robot to assist children with ASD	bodied humanoid robot	interviews with the users
Probo [36]	social robot for children with ASD	compliant actuation systems	children performance
Healthcare CAs			

CA	Short description	Main technology	Evaluation Method
CoachAI [89]	patient's support chatbot	task-oriented finite state machine (FSM) architecture	user's engagement, system accaptance and rating.
Woebot [96]	therapist CA	AI,NLP,empathy engine	users' reports
Mandy [202]	a primary care CA	NLU,NLG,word2vec	accuracy
Tanya [85]	graphically embodied female agent that supports breastfeeding		increased breastfeeding success
KR-DS [301]	diagnosis chatbot	Bi-LSTM, Deep Q-network	diagnosis accuracy
Commercial CAs			
CA	**Short description**	**Main technology**	**Evaluation Method**
SuperAgent [71]	customer service chatbot	AIML+LSA	2 customer reviews
SamBot [225]	questions answering CA	AIML	Loebner Prize Competition
			+ user interaction

Finally, Fig. 10 illustrates the various evaluation methods and their relation to each of the relevant components.

3.7 Publicly Available Conversation Datasets

Conversation datasets are used to train machine learning CA models and to test the quality of the CA. In this section some of the existing datasets used in the literature for CA development and CA evaluation are described. Some recent reviews focusing on available conversation datasets are presented next.

Serban et al. [259] review different types of conversations datasets for CAs and categorize them according to the type (text or speech), topics, length (number of dialogs, average number of turns, and number of words), and description.

Keneshloo et al. [139] provide a list of conversational datasets that can be used for sequence-to-sequence models. Some of the databases provided can be helpful for the dialogues

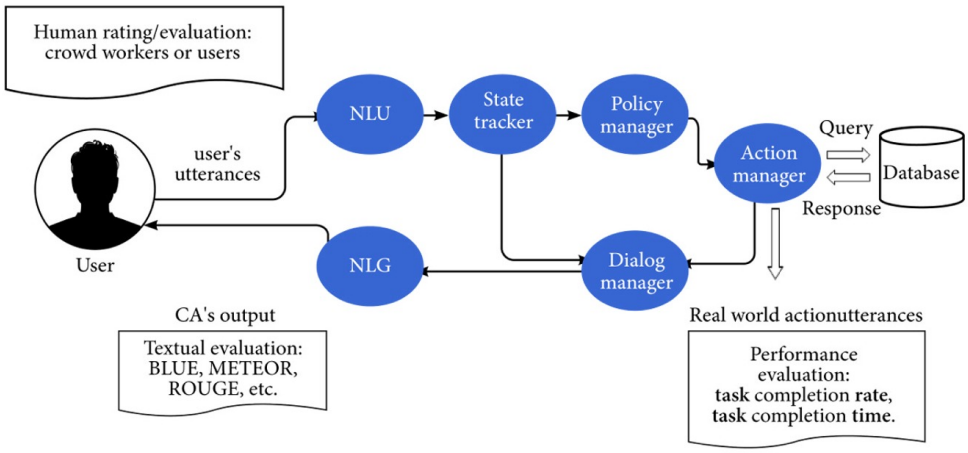

Figure 10: A diagram illustrating the various CA evaluation methods.

generated by conversational agents, and others are related to other domains, such as image and video captioning, computer vision, speech recognition, and synthesis.

Deriu et al. [77] provide another list of available conversation corpora focusing on task related conversations in several domains, such as the restaurant domain and the tourist information domain. They note that question answering dialog systems can be extracted either from chat logs or from several available literature sources, news, scientific resources, Wikipedia articles, FAQ sites, and even cooking domains.

In the remainder of this section, some of the most useful corpora for conversation understanding, generation, and evaluation are described and classified according to their applications, using the terms defined in Section 3.1.

Datasets for General Purpose CAs

There are various sources of datasets used for general purpose dialogues. DailyDialog[4] [163] is a dataset consisting of handwritten texts, manually labeled with communication intention and emotion information. DailyDialog contains multi-turn dialogues, reflecting daily communication on various aspects of daily life. The dialogues in the dataset conform

[4] `http://yanran.li/dailydialog`

to various common dialog flows, such as question and answer, bi-turn flows, and multi-turn dialog flow patterns reflecting realistic dialogs.

Large amounts of available data on movie reports may also be utilized to build dialogue corpora. The SubTle corpus [21] is designed for general purpose interaction generation. It is composed of interaction-response pairs, extracted from the OpenSubtitles[5] [168, 278] movie corpus, which is a multi-language conversation corpus based on movie subtitles. Additional datasets based on movie dialogs are the Movie dialog dataset[6] [82] and Cornell movie dialogues corpus[7] [73].

Serban et al. [259] consider the advantages and disadvantages of training and evaluating CAs based on artificial datasets, such as datasets extracted from movie manuscripts and audio subtitles. The advantages are as follows: (a) the dialogues resemble human spontaneous language; (b) the dialogues are easy to follow and contain less garbling and repetition; (c) there is a diversity of dialogues, topics, environments, actors, and relationships. This enables creating a more flexible CA, which may talk with various users in different situations while using various interaction patterns. However, since CAs must consider the context to provide accurate responses, Serban et al. state that artificial datasets may have a caveat as they do not provide this context. It should be noted that since dialogues from movies can be too extreme and not reflect real-life dialogues, training and evaluating CAs based on them may lead to undesired behavior on the part of the CAs.

Another source of datasets, for the training and evaluation of CAs, is the social media. Many datasets are composed of texts extracted from popular conversation websites and applications, such as Reddit[8] and Twitter[9].

Dialogue corpora based on Twitter conversations are developed and used by Li et al. [158], Sordoni et al. [265], Xu et al. [300], and Ritter et al. [242]. Dialogue corpora based on Reddit forums have been developed by several other studies, including the study of Dodge et al. [82], Serban et al. [257], Schrading et al. [254], and recently by Zhang et al. [316]. The dialogue generation model of PLATO [39] is pretrained on both Twitter and Reddit. The Ubuntu dialogue corpus [175] is based on the Ubuntu chat logs.

Serban et al. [259] note that datasets based on conversations extracted from social media, have some significant limitations. Generally, they are noisy, and they may include texts generated by non-human CAs, such as influence agents. Another limitation of Twitter-based datasets is the maximum length of 140 characters per Twitter message. As a result, the

[5] http://opus.nlpl.eu

[6] https://www.kaggle.com/abhishek/the-movie-dialog-dataset

[7] https://www.cs.cornell.edu//~cristian/Cornell_Movie-Dialogs_Corpus.html

[8] https://www.reddit.com

[9] https://twitter.com

Twitter corpus has an enormous number of typos, slang, and abbreviations as well as Twitter-specific structures, such as hashtags. Similar to the issue with artificial datasets, Serben et al. note that dialogues extracted from social media may be missing context. In addition, as stated by Kourosh [18], the use of auto-correction by users of social media may cause an additional layer of complication.

Datasets for Question Answering

Question answering conversational agents can be trained using publicly available question and answer web pages. Zeng et al. [313] survey machine reading comprehension evaluation and benchmark datasets. They note that the most popular datasets in this category are the Stanford question answering dataset (Squad) versions 1.1 [233] and 2 [232], the CNN/daily mail dataset [120], the natural questions dataset [149], and TriviaQA [134].

The Squad datasets are designed for machine reading comprehension training. They consist of more than 100K questions and answers posed by crowd workers in Wikipedia articles; the answers are citations within Wikipedia articles. The CNN/daily mail dataset contains question/answer pairs generated from CNN and daily articles, published during 2007-2015 for CNN and during 2010-2015 for the daily mail.

The natural questions dataset [149] contains real user questions posted on Google search, and answers found on Wikipedia by crowd workers. Each real question may have three types of answers: an associated long answer, which is based on text from a Wikipedia article, a list of short answers, and a yes-no-answer.

Finally, the TriviaQA [134] dataset, designed for machine reading comprehension challenges, contains triplets of question-answer-evidence; the evidence aims to ease the answering process. TriviaQA contains relatively complex and challenging questions with syntactic and lexical variability, requiring cross sentence reasoning in answering TriviaQA questions.

Datasets for Goal Oriented CAs

The challenge of designing a goal-oriented CA is twofold: the CA should be both effective in NLU and NLG, and efficient in helping to solve the common task. Consequently, the task-oriented conversation should take into consideration both aspects. A useful source for obtaining goal-oriented datasets is the dialog system technology challenge (DSTC) [119], which is a yearly challenge started in 2013. Various well-known datasets have been produced and released for every DSTC edition.

The schema guided dialogue (SGD) dataset [237], released for DSTC8, contains approxi-

mately 23K annotated multi-domain (bank, media, calendar, travel, weather), task-oriented dialogues between a human and a virtual assistant. SGD can test state tracking as well as intent prediction, slot filling, and language generation.

MultiWOZ [54] is a tourist dialog dataset, annotated with dialogue belief states and dialogue actions. The dialogues in MultiWoz cover seven touristic domains: attractions, hospitals, police, hotels, restaurants, taxis, and trains. Each dialogue in MultiWoz can cover more than one domain.

Taskmaster-1 [56] includes dialogues of the following task-oriented domains: ordering pizza, setting auto repair appointments, arranging taxi services, ordering movie tickets, ordering coffee drinks and making restaurant reservations. More than half of the dialogues were created manually, using crowd-workers to compose entire dialogues.

Finally, MultiDoGo [223] is a public human-generated multi-domain dialogue dataset, composed of dialogues created by crowd workers and trained annotators, with a total of over 81K dialogues across six domains. Over 54K of these conversations are annotated for intent classes and slot labels.

For a list of task-related datasets, including DTSC challenges datasets, see Deriu et al. [77].

Datasets for Social Assistance

Social assistance CAs aim to provide medical, healthcare, mental, or other educational assistance. In these domains, there may exist a privacy issue: information in medical, mental, or educational dialogues is sensitive, and therefore, it is difficult to publish dialogues in a way that would honor the privacy of the participants. Here are some repositories found in these areas.

The first attempt to create a large medical corpus is MedDialog, developed by Zeng et al. [314]. MedDialog is a medical dialogue dataset that consists of 3.4M conversations between patients and doctors in Chinese, covering 172 specialties of diseases, and 260K conversations in English, covering 96 specialties of diseases. Each consultation consists of a description of the patient's medical condition, followed by a conversation between the patient and the doctor. The data are gathered from Iclinic[10] and HealthcareMagic[11], which are online health care service platforms.

Another health-related dataset was constructed by Yang et al. [304]. Their dataset consists of a collection of conversations in English and Chinese between doctors and patients about COVID-19. The English dataset contains 603 consultations, and the Chinese dataset contains

[10] `iclinic.com`

[11] `caremagic.com`

1088 consultations.

Sharma et al. [260] introduced the task of transforming low-empathy conversational posts into higher empathy posts. They focus on mental health-related conversations, filtered from posts of TalkLife[12], which is the largest online peer-to-peer support platform for mental health support. The dataset contains 3.33M interactions from 1.48M users posts. The interactions were labeled with empathy measurements using a framework, consisting of three empathy communication mechanisms: emotional reactions (expressing emotions such as warmth, compassion), interpretations (communicating an understanding feelings and experiences), and explorations (improving understanding of the users by exploring feelings and experiences).

Another dataset that can be used for empathic user responses is EmpatheticDialogues[13] [236]. This dataset consists of 25K conversations grounded in emotional situations, divided into 32 different emotion categories. The conversations are open-domain and handled between two users, with one responding empathetically to the other. Next, some datasets are described that may be helpful in recognizing emotion, detecting abuse, and generating empathic responses, which are all qualities expected from a CA used for mental and psychological assistance. The emotionally recorded corpus SEMAINE, developed by McKeown et al. [184], is based on recorded dialogues of users talking with an operator who tries to evoke emotional reactions. The corpus includes 20 participants and 100 conversations, all recorded with high resolution cameras and microphones.

Schrading et al. [254] built a text dataset of domestic abuse, extracted from Reddit. The dataset includes abuse and non-abuse texts. Chai et al. [63] developed an offensive response dataset, which consists of 110K input-response chat records in which the response is either appropriate or offensive. These databases can assist in training CAs, allowing the CAs to identify different sensitive situations to respond accordingly.

Educational Datasets

Here, educational datasets that can be helpful for educational CA development are provided.

The BURCHAK dataset [311] is a human-human dialogue dataset for interactive learning of visually grounded word meanings in a foreign language. A learner needs to learn invented words for visual objects (for example, the word "burchak"for a square) from a tutor. The text-based interactions resemble face-to-face conversations and thus, contain many of the linguistic phenomena encountered in spontaneous dialogues. The corpus contains 177 conversations and

[12] `talklife.co`

[13] https://github.com/facebookresearch/EmpatheticDialogues

includes 2454 turns in total.

Wolska et al. [297] annotate a corpus of tutorial dialogues on mathematical theorem proving. To collect the data, they design and perform an experiment with a simulated tutorial dialog system to teach mathematical theorem proofs. The total corpus comprises 66 sets of dialog session logs with 12 turns, on average. There are 1115 sentences in total, of which 393 are student sentences.

Hutzler et al. [128] prepared a bank of questions designed to train high school students on reading comprehension skills. The questions were rated by a panel of experts using a set of criteria based on Bloom's cognitive taxonomy [47].

The CIMA collection [266] includes tutoring dialogues between crowd workers playing the role of students and tutors. The tutoring utterances include educational strategies, such as hint provision and questions asked to check the student's understanding.

MyPersonality[14] is a knowledge base composed of information collected from over 6 million volunteers on Facebook using a personality questionnaire. MyPersonality is used by KBot [14], a social media trained chatbot, to find answers to some questions that cannot be found in other knowledge bases, especially in the psychological and social science domains.

Tables 4 and 5 describe the list of datasets available online, which are reviewed in this section. For each dataset, a short description is provided along with some important attributes and the type of conversational agent that uses it, referring to the usage described in Fig. 3.

[14] `http://mypersonality.org`

Table 4: Main available datasets for conversational agents - part A

General Purpose Datasets				
Dataset	**Source**	**Description**	**Size**	**Used for**
DailyDialog[15] [163]	handly written, manualy labeled	daily interactions	13,118 dialogs, $\tilde{7}$.9 turns	general purpose
[278]	subtitles	interaction-response pairs		general purpose
Movie Dialog dataset[16] [82]	movie metadata as knowledge triples	OMDb, MovieLens and Reddit	3.1M simulated QA pairs	Movies QA and recommendation
Cornell Movie Dialogues Corpus[17][73]	Short conversations from film scripts	movie metadata	220K conversations	understanding linguistic style
Ubuntu dialogue corpus[18] [175]	Ubuntu chat stream	human-human chat	930K conversations	response generation
Question Answering Datasets				
Squad Version 1.1[19] [233]	question& Answers on Wikipedia articles	$\tilde{1}$00K questions on Wikipedia articles	100K q&a	machine reading comprehension
Squad Version 2[20] [232]	question and Answers and additional questions with no answers	Squad 1.1 + 50k questions with no answers	100K Q&A + 50k questions	machine reading comprehension

[15] http://yanran.li/dailydialog
[16] https://www.kaggle.com/abhishek/the-movie-dialog-dataset
[17] https://www.cs.cornell.edu//~cristian/Cornell_Movie-Dialogs_Corpus.html
[18] https://github.com/rkadlec/ubuntu-ranking-dataset-creator
[19] https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/
[20] https://rajpurkar.github.io/SQuAD-explorer/

CNN/Daily Mail[21] comprehension [120]	queries from the CNN and Daily Mail websites	cont.–query–answer triples	~1M stories+ associated queries	machine reading training dataset
Natural Questions [22] dataset [149]	Google search queries+ Wikipedia answers by crowd workers	Google question+ long answer+ short answers	307,372 training examples	training & evaluation of answ. systems
TriviaQA[23] [134]	crowdworkers questions	question-answer-evidence triples	95K quest.-ans. pairs + 6 evidence doc. per quest.	reading comprehension

[21] https://github.com/deepmind/rc-data

[22] https://github.com/google-research-datasets/natural-questions

[23] http://nlp.cs.washington.edu/triviaqa/

Table 5: Main available datasets for conversational agents - part B

Datasets for Goal Oriented CAs				
Schema Guided [24] Dialogue [237]	dialogue simulator+ paid crowd-workers	multi-domain, task-oriented human-agent convev.	20k conversations	intent prediction, lang. generation, dialogue tracking
MultiWOZ[25] [54]	turkers working	human-human	10k dialogues conversations	Task-oriented dialogue modelling
Taskmaster-1[26] [56]	crowd workers users and center operators	spoken & written technical dialogs	5,507 spoken & 7,708 written dialogs	dialog systems research, dev. and design
MultiDoGo[27] [223]	crowd workers paired with trained annotators	human to human, services dialogues	$\tilde{8}$1K dialogues across 6 domains,	virtual assistants development
Datasts for Supporting CAs				
Covid-19 dialogue[28] dataset[304]	online health care platform[29]	conversations between doctors and patients	603 Eng. + 1088 Chinese consultations	medical dialogue system systems
MedDialog[30] [314]	medical dialogue platform[31]	doctors-patients conversations	1.1M Chinese + 0.3M English dialogues	medical dialogue systems

[24] https://github.com/google-research-datasets/dstc8-schema-guided-dialogue
[25] https://github.com/budzianowski/multiwoz
[26] https://g.co/dataset/taskmaster-1
[27] https://github.com/awslabs/multi-domain-goal-oriented-dialogues-dataset
[28] https://github.com/UCSD-AI4H/COVID-Dialogue
[29] haodf.com
[30] https://github.com/UCSD-AI4H/Medical-Dialogue-System
[31] haodf.com

SEMAINE [184] [32]	human-human conversation experiment	emotionally coloured conversations video recordings	25 recordings,$\tilde{3}$0 minutes long	eliciting non-verbal signals in human-computer interactions
Empathetic-[33] Dialogues[236]	810 crowd workers select an emotion and talk about it	conversations grounded in emotional situations	25k conversations	recognizing human's feelings
Offensive [34] response dataset [63]	input-response records from SimSimi[35], offensivity annotated by crowd workers	input-response pairs and their annotation	110K chat pairs	improve CA abilities
BURCHAK [36] dataset[311]	dialogues of pairs of participants, discussing visual attributes of 9 objects	chat outputs of dialogues	177 dialogues 2454 turns	learning visually grounded word meanings in a foreign language
The CIMA collection[37] [266]	conversations between crowd workers playing as students and tutors.	tutoring interactions and accompanying responses	2,970 tutor responses to 350 exercises.	tutoring conversation based on a provided strategy.

[32] https://semaine-db.eu/
[33] https://github.com/facebookresearch/EmpatheticDialogues
[34] https://github.com/chaiyixuan/Offensive-Responses-Dataset
[35] www.simsimi.com
[36] https://sites.google.com/site/hwinteractionlab/babble
[37] https://github.com/kstats/CIMA

4 Detecting Harmful and Insulting Situations Via Text

Given the extensive scientific background on conversation agents, we would like to continue the development of an agent that can advise the child with special needs how to respond and what to do in different social situations. For this purpose, we will develop a system that will identify situations that require intervention according to the text that the child in the child's conversation by him or by the person talking to him. In order to help these children, the agent must be aware of the child's interactions, translate the audio contents into text, recognize the text classification, and detect if a special situation occurs (i.e.a risky situation, or a situation involving insulting context). By analyzing all the information, the agent will advise the user of the proper behavior in that case. For example, if someone unfamiliar tells the child, "Let's get in the car and I'll give you a lollipop", the agent can detect a dangerous situation and recommend that the child escapes. The agent may also alert the child's guardians. Another example may be when the child tells his grandmother: "You're old" the agent will recommend that the child says something positive to her grandmother, such as "You have a lot of life experience and I love to hear your life story!".

4.1 Dataset Details

In the previous section we described some related work concerning sentiment classification. Most of the datasets used in previous related studies are based on comments about movies or services (e.g., movie review) or on forums or twitter posts. However, the text said by people or children may be different than such reviews or posts, because talking at home, in class or near friends, etc., can be different from the terms used in written text such as comments or forums. This is especially true when considering children's conversations. Consequently, insulting context as well as language indicating threats may be different. Given this difference, available on-line dataset resources of essays, comments and recommendations are not entirely appropriate for training an agent in determining types of spoken conversation.

Our dataset for the text analyzing stage was built as following: The sources of the sentences were taken from an initial seed of 100 unintentional insulting sentences obtained by performing interviews with parents of children with ASD, performed by the autism center, as described by [187] and another group of sentences provided by workers of MTurk [20], in response to our surveys: The following survey (HIT) was run using Mturk [20], to construct sentences for the conversation database:

We are conducting a research and development agent designed to help children with special needs, analyze their environment and help them respond correctly to the social situations they face.

1. Sometimes a child might say sentences that insult the listener, For example: "Grandma, you are fat."
Please provide 5 examples of such sentences.

2. There are sentences that can insult the listener, but only in certain situations. For example: "When was the last time you straightened up your room?" "Are you sure you know what you are doing?"
Please provide 5 examples of such sentences.

3. Some sentences spoken by the children may not be related to the previous discourse at all, or are repetitive or strange. For example: "Exactly 654 seconds ago, Father left for work."
Please provide 5 examples of such sentences.

4. Children with special needs may be at risk of various types of abuse and bullying. We want to identify sentences that children are told and may indicate that they are at risk. For example: "Do not tell anyone our secret..." or: "Come here, I want to give you a hug."
Please provide 5+ examples of such sentences.

And in an additional survey, we asked for examples of the fourth type of sentence only, namely bullying and at risk and using the same illustration as above we asked for 10 examples of such sentences.

In this manner the MTurk workers, who were located in USA, were asked to provide sentences for each of the following categories: insulting sentences, sentences which are context dependent, repetitive or strange sentences (which were associated to the non-insulting sentences), and sentences indicating risk. The payment per assignment was 1$, and we collected 83 assignments. In order to increase the number of sentences that indicated risk, we performed an additional survey explained above and they received 0.1$ for each assignment. In the second survey we collected 51 such assignments and approximately 2170 sentences were gathered in this manner.

Some of the sentences were collected offline by students who were asked to provide relevant sentences. Additional sentences were gathered from expert talks about safety, and in particular, safety of children with special needs. Another source was text from on-line groups and forums, concentrating on groups of children with special needs. Furthermore, other sentences were taken from news articles and from responses to news articles, where we collected

Table 6: Distribution of Sentence Types

Sentence Type	Count	Frequency
Normal sentences	2910	21.6%
Context-dependent[38]	2269	16.8%
Insulting third person	2644	19.6%
Insulting sentences	3511	25.9%
Sentences indicating risk	2173	16.1%

sentences that can be said by children, or to a child.

Our dataset contains context relevant to children, and it is categorized into five categories: Nonetheless, in the current study we did not consider the context of dependent sentences, since deciding about them requires additional information about the situation, rather than the text of the sentence itself.

The distribution of the sentence types is described in Table 6.

Typically, different types of sentences have different sets of common words. Figure 11 presents the frequency of the common words for the types considered in this study. As we can see, the typical common words are different for different types of sentences, though there are words that appear frequently in different types of sentences. For example, the word *you* is the most common word in most of the types, except in the third person insulting type, where the most common word is 'is'.

Sentence Type	**Sentence length (min..max length)**	**Average len and (std)**	**Vocab. size**
Normal sentences	1..36	6.5 (3.02)	2,611
Context dependent	1..26	6.95 (2.77)	1,957
Insulting third person	1..46	7.21 (3.33)	2,746
Insulting sentences	2..31	6.92 (3.12)	2,855
Sentences indicating risk	1..23	7.78 (3.23)	1,467

Table 7: Distribution of Sentence length

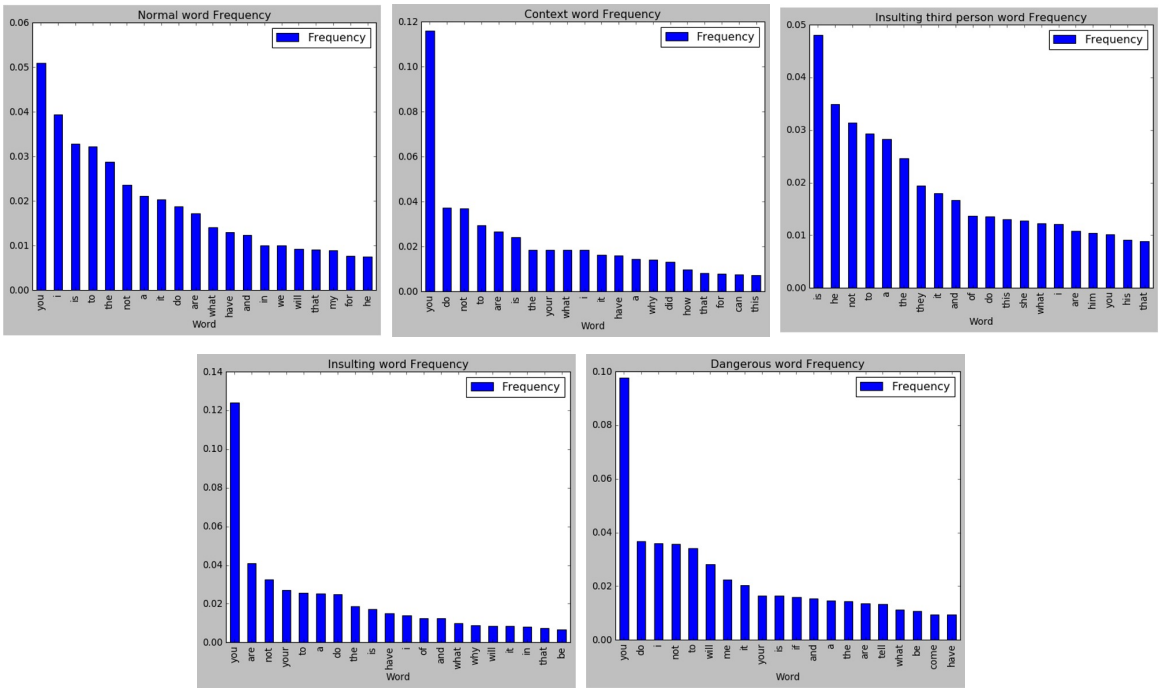

Figure 11: Common Word Frequencies

4.2 Pre-processing of Text Dataset

The details of the sentence lengths in the database (DB) are described in Table 7. As depicted, the average length of the sentences is very similar in the different sentence types.

In order to prepare our dataset for the different methods, we ran two different pre-processing algorithms. Since the categorization algorithms we used belong to two groups: (a) classical machine learning algorithms implemented by the Scikit-learn Python library; (b) the Embedding-CNN method implemented by Keras, using the TensorFlow backend, we used a different pre-processing algorithm for each of these groups, i.e., Algorithm 1 for the Scikit-learn based methods,and Algorithm 2 for the Embdedding-CNN method. Algorithm 1 runs some generalizations on the word of the sentences (from both the training set and the test set), then a bag-of-words is created for each sentence, transformed to TD-IDF,

(TF-IDF measures the importance of a word in a document, in the context of a collection of documents containing that word. Words that appear frequently in all documents within a corpus are considered differently than those that appear frequently in just one of the documents.)

which is sent to the machine learning algorithm. Algorithm 2 starts with some more simple preprocessing of the sentences, while adding relevant phrases from the movie review dataset [214] to both the training set and test set, as described below. then it uses Word2Vec based

on the Google news vector [104] for the embedding process, and the result of the Embedding process is sent to the CNN. These algorithms were developed especially for our work.

Algorithm 1. Preprocessing Sentences

Input: Sentence Datasets DS
Choose 2100 sentences for each category $\in \{NS, ITP, ISP, RSK\}$
where $NS = normal-sentences$,$ITP = insultingthirdperson$, $ISP = insultingsecondperson$, $RSK = sentencesindicatingrisks.$
Create a list $Negative$ *for negative terms, a list* $Positive$ *for positive terms.*
Read the dictionary $CategoriesDictionary$ *of* $[word, categories]$
For each sentence $S \in DS$*:*
    Lemmatize S*, using* $WordNetLemmatizer$
    If first word of S *is* $\in \{'is','are','be','do','did','have'\}$,
        Replace it with $questif.$
    If a word $w \in \{'is','are','be','do','did','have'\}$ *appears in* S*, following by not,*
        remove w
    Replace the term $\{'havenot','neednot'\}$ *with must.*
    Remove each sign $\in \{'.','\,','\,','\,';','\,a','\,!','<','>','\,...','\,?'\}$
    $toAdd = \{\}$
    For each $word \in S$*:*
        If $word \in Positive$
            $toAdd = toAdd|\{positive\}$
     Else $if word \in Negative$
        $toAdd = toAdd|\{negative\}$
    For each $word \in toAdd$
        concate $word$ *to* S
    for each $word \in S$
      If $word \in CategoriesDictionary$
        replace $word$ *by* $CategoriesDictionary[word]$
Randomly split DB *to* 90% *training-set,* 10% *test-set.*
For each sentence in DB
    Put sentence into a bag-of-words form
    transform the sentence representation to TD-IDF form
return $transformed(trainingset), transformed(testset)$

Next, Algorithm 2 provides the preprocessing details for the Embedding-CNN learning method. Note that steps 5-9 in Algorithm 2 were developed by Wang et al. [291].

Algorithm 2. Preprocessing for Embedding-CNN learning method:

Input: Sentences Datasets DS*, where* $DS = \{(sentence, type(sentence)|sentence \in SetOfSentences\}$

Create a list $Negative$ *for negative terms, a list* $Positive$ *for positive terms.*
Read a dictionary $CategoriesDictionary$ *of* $[word, categories]$
For each sentence $S \in DS$*:*
    Run a lemmatization process over S*, using* $WordNetLemmatizer$
    $toAdd = \{\}$
    For each $word \in S$*:*
        If $word \in Positive$
            $toAdd = toAdd|\{positive\}$
        Else $if word \in Negative$
            $toAdd = toAdd|\{positive\}$
    For each $word \in toAdd$
        concate $word$ *to* S
$phraseDictionary = \{\}$, $phrasesList = []$
For each $(sentence, type(sentence)) \in DB$
    For each $phrase = word_i..word_j \in sentence$*:*
        $phraseDictionary[phrase] = phraseDictionary[phrase] + [type(sentence)]$
For each $phrase \in phraseDictionary$*:*
    Find the common type and the percentage of phrase in $phraseDictionary[phrase]$
    If (phrase contain one word Or phrase appears 10 times or more in DB) and the percentage of the common type is greater then or equal to 0.75)
        Add phrase to $phrasesList$
    For each $p \in MoviewReviewDB$
        Add p *to* $phrasesList$
Randmoly split DB *to* 90% $training - set$ *and* 90% $test - set$
For each $phrase \in phrasesList$*:*
    If phrase appears in the text of $training - set$ *and phrase does not appear in the text of* $test - set$*:*
        Add phrase to $training - set$
return an embedded vector of $training - set, test - set$*, using the embedding data of GoogleNews-vectors-negative300.bin [104]*

It is noteworthy to emphasize that the training set built in the preprocessing algorithm of the embedding+CNN method includes, in addition to 90% of the original sentences, also additional sentences that were used for the training set. Some of the sentences, with a clear meaning (appear 10 times or more in the database, with a frequency of 75% or more of appearing in one of the types), were added to the training set as phrases, if they were not substrings of the sentences of the test set. In addition, phrases taken from the Movie Review (MR) database that were not sub-strings of the test set were also added to the training set.

4.3 Methodology Description

The first method we used was the **Extra-Tree** method (which stands for extremely randomized trees) that was proposed in [211], with the main objective of further randomizing tree building in the context of numerical input features. In this case, the choice of the optimal cut-point is responsible for a large proportion of the variance of the induced tree.

The most successful method was the Random Forests.

Random Forests are bagged decision tree models. Each decision tree in the forest considers a random subset of features when forming questions and only has access to a random set of the training data points. This increases diversity in the forest leading to more robust overall predictions and the name 'random forest'. In our study, the Random forest was based on 100 estimators, and as described below in Section 4.4, it reached the best results.

KNN algorithm is a robust and versatile classifier that is often used as a benchmark for more complex classifiers such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Despite its simplicity, KNN can perform better than more powerful classifiers and is used in a variety of applications such as economic forecasting, data compression and genetics.

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm outputs an optimal hyperplane, which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane into two parts where each class lies on a different side of the hyperplane.

Ridge Classifier works similarly to LogisticRegression with a l2 penalty, but it uses the Ridge regression model for multi-class classification in the following way to create a classifier: 1.Use a label binarizer to create multi-output regression, one for each class (One-Vs-Rest modelling) and train the Ridge regression model. 2.Get a prediction from each class' Ridge regression model (a real number for each class) and then use argmax to predict the class.

The **Naive Bayesian** classifier is based on Bayes' theorem with the conditionally independent assumptions between features. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

MultiLayer classifier implements a multi-layer perceptron (**MLP**) algorithm (a neural network). MLP is a supervised learning algorithm that learns a function by training on a dataset. Given a set and a target, it learns a non-linear function approximation for either classification or regression. It is different from logistic regression, because there can be one or more non-linear layers, called hidden layers, between the input and the output layers.

MLP trains on two arrays: array X of size (n-samples, n-features), which holds the training samples represented as floating point feature vectors; and array y of size (n-samples), which holds the target values (class labels) for the training samples. We used a network with three hidden layers, each containing 100 sigmoid nodes.

Voting is a classifier that trains all the above methods, and then for each sentence of the test set, performs a voting protocol over the above methods and chooses the category suggested by the majority. The methods used in the Voting classifier are: Random forest, Extra trees, KNN, SVM, Ridge Classifier, Bayesian inference method, and MLP.

Next, we describe the template of the Convolutional Neural Network (CNN) used for our text classification task. CNN is a class of deep, feed-forward artificial neural networks (where there are no cycle connections between the nodes) that use a variation of multilayer perceptrons designed to utilize minimal preprocessing. These are inspired by animal visual cortex.

In CNN the result of each convolution will dismiss when a special pattern is detected. By changing the size of the kernels and concatenating their outputs, allows the detection of patterns of variant sizes (2, 3, or 5 adjacent words). Patterns could be expressions (word ngrams) like "I hate", "very good" and therefore CNNs can identify them in the sentence notetheless to their position.

The structure of the CNN used is taken from [178], where a CNN template for classification is suggested, and their template reached the best result for our database. In this model, the first convolution layer used had a filter length of 5 and ReLU as its activation function. The next part is a maxPooling layer, followed by a dropout of 0.2. followed by two additional convolutional and maxPooling layers,then a simple layer with 128 neurons and a ReLU activation function, and finally, a softmax layer with five outputs, one for each category.

4.4 Experimental Results

First, we describe our results using classical machine learning methods, imported from the Scikit-learn library, on the sentence databases, preprocessed using Algorithm 1. We ran 50 experiments, where in each of them, the sentence dataset was randomly split into a training set and a test set. Table 8 shows our results.

As depicted in the table, the extra trees method achieved the best results, with the ability to correctly predict the type (normal sentence, insulting sentence, third person insulting sentence, or sentence indicating risk) with 71% accuracy and an F1 score of 0.710. Other successful methods, with very close performance, are the random forests (with 70% accuracy and 0.702 F1 score) and Ridge Classifier (with 67.8% accuracy and 0.672 F1 score). The

method	**average and (std) accuracy**	**average and (std) F1 score**	**average and (std) precision**	**average and (std) recall**
Random Forests	0.703 (0.013)	0.702 (0.013)	0.702 (0.013)	0.704 (0.013)
Extra Trees	0.711 (0.015)	0.710 (0.015)	0.711 (0.015)	0.712 (0.015)
KNeighbors	0.549 (0.016)	0.545 (0.016)	0.551 (0.016)	0.549 (0.015)
SVM	0.678 (0.014)	0.672 (0.014)	0.678 (0.015)	0.679 (0.014)
Ridge Classifier	0.680 (0.015)	0.678 (0.015)	0.679 (0.015)	0.680 (0.015)
Bayes	0.628 (0.016)	0.626 (0.016)	0.635 (0.016)	0.628 (0.016)
MultiLayer	0.643 (0.017)	0.643 (0.017)	0.645 (0.017)	0.643 (0.017)
Voting	0.711 (0.016)	0.711 (0.016)	0.711 (0.016)	0.712 (0.016)

Table 8: Accuracy Results

Voting classifier reached solutions very close but slightly higher than that of the Extra Trees method (average accuracy of 71.1% and average F1-score 0.711). The confusion matix are in A We will now continue with a description of our results from the Embedding-CNN method. This method first runs the preprocessing algorithm described in Algorithm 2, and then applies the CNN method described in Section 4.3. We used a batch size of 128, trained the network for 10 epochs, and used the Adam optimizer [144].

After running 50 runs, the average accuracy level reached by the CNN on the test set was 71.05% (std 0.0067) and the F1 score was 0.70 (std 0.0067) the precision was 0.713 (std 0.009) and recall was 0.695 (std 0.006) . Note that, as described in Section 4.2, the embedding-CNN method was trained on 90% of our conversation database, and in addition, phrases from movie reviews that were also used in the training set. With this combined training set, the accuracy of the CNN was higher than most of the machine learning methods, Nonetheless, a random forest method, with 100 estimators, and the Voting classifier, reached slightly higher results, while it required a smaller training set and shorter training time w.r.t. the Embedding-CNN method.

Finally, we checked whether a set of neural networks can achieve better results than a single network. Thus, we created a random generated panel of 10 CNN based classifiers. The structure of each classifier was as follows: after the embedding process, a 1-D convolutional level was used, with 128 filters and a softplus activation function. Then, a max pooling process was performed, followed by a dropout of 10%. Then, another 1-D convolutional layer was used with 32 filters and a linear activation function followed by max pooling. Thereafter, a third 1-D convolutional layer was used, with 128 filters and a hyperbolic tangent activation

method + max pooling. Then, a flatten layer (size 128) with a sigmoid activation function was used, and its outputs were sent to a softmax layer. The batch size was set to 64, and we ran 10 epochs. Each classifier was trained on 90% of the training set, and we chose the best five classifiers, based on their accuracy on the validation set (the remaining 10% of the training set). We then, determined the type of each sentence by a vote between the five best classifiers, which we called the panel. This voting panel increased the accuracy and the F1 score of the classifications. In particular, after 50 runs, the average accuracy rose to 72.2% (std 0.009) with an F1 score of 0.714 (std 0.009), resulting in higher accuracy and F1 scores than that reached by each of the experts individually.

4.5 State of the Art Methods for Text Emotion Recognition - A Comparison

The first step in our work was text emotion recognition, a very common task for which several state-of-the-art methods have already been developed. We ran some of these on our dataset.

XLNet [306] is a generalized autoregressive pretraining method. It enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order. Yang et al. build XLNet on a generalized autoregressive method that leverages the best of both autoregressive language modeling and autoencoding language modeling while avoiding their limitations. They used the BooksCorpus and English Wikipedia as part of their pretraining data. To aggressively filter short or low quality articles for ClueWeb 2012-B and Crawl Common they used heuristics. After tokenization with SentencePiece, they obtained 2.78B, 1.09B, 4.75B, 4.30B, and 19.97B subword pieces for Wikipedia, BooksCorpus, Giga5, ClueWeb, and Common Crawl respectively. They tested XLNet on the Yelp5 corpus and achieved a best result of 73% accurracy. We took the XLNet system and ran it on our dataset, following their protocol: We divided our dataset, using 90% for training, and 10% for testing. We ran the training set 4,000 times.After every 1000 training runs, we ran our test dataset and saved that checkpoint. Finally,after comparing these four checkpoints, we tested the most successful of them by running it on our dataset with four classifiers and also got a best result of 73% accruracy.

Next, we used a Convolutional Neural Network (CNN) for our text classification task. CNN is a class of deep, feed-forward artificial neural networks (where there are no cycle connections between the nodes) which use a variation of multilayer perceptrons designed to utilize minimal preprocessing. These are inspired by the visual cortex of animals. We used the CNN model developed by Maheshwari et al. [178] and started with Google Glove 6B vector 100d, an unsupervised learning algorithm for obtaining vector representations for words.Then

we used a very simple convolutional architecture, using a total of 128 filters with size 5 and max pooling of 5 and 35. The CNN was trained for 10 epochs, using Adam optimizer [144], with a batch size of 128. The average accuracy level on the test set was 69.6% (std 0.008) and the F1 score was 0.681 (std 0.008).

The CNN based method will normally beat the best-performing classical ML methods, but here that was not the case. A possible explanation for this is our dataset, which consists of relatively short sentences (up to 500 characters). The power of CNN is in its ability to extract feature patterns from an image or from long sequences, where these features can appear in several places. However, in our datasets, where the input size is relatively limited, the strength of the CNN method is not expressed.

5 Embedded Vectors for Detecting Harmful and Insulting Situation Via Text and Voice

After describing our results on text, we wanted to test whether a combination of voice media would improve the detection results of different risks and threats. Therefore, in the second stage we combined both text and voice in order to improve our results and proceed to our real world goal.

5.1 Dataset Details

At the second stage we built another dataset which was composed from text and voice sentences: The dataset is composed of 2677 Hebrew sentences, including both text and audio. These sentences were extracted by manually splitting YouTube videos and open children's series into sentences. We took the various YouTube videos from different sites that deal with teaching complex social situations relevant behavior, as well as parents who shared different recordings and news sites documenting different dialogues between people. Finally, the sentences were classified by a team of educators, into three categories: neutral speech (900), insulting speech (963), and unsafe speech (814).

Figure 12 shows the distribution of the three categories in the dataset. All software and data will be made available.

We analyze the common words for each of the categories in the dataset. Figures 13-15 present the frequency of the 20 most common words for each category. As can be observed, the common words are different for each category. In all tree types of sentences we investigate only words of $length > 3$.

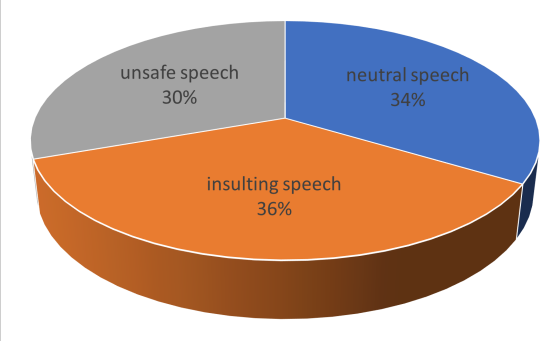

Figure 12: Dataset distribution into categories

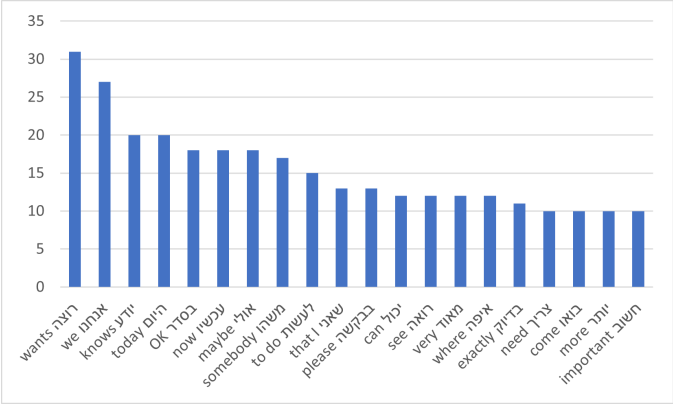

Figure 13: Hebrew Neutral Sentences Word's Frequency

The most common words in the neutral category are the Hebrew words meaning: 'wants', 'we', 'OK', 'today', 'knows', 'where', and 'please'.

For the insulting category, the most used words were: 'at all','wants', 'more', 'me','which', 'to be', 'that he' and 'need'. This may indicate that insulting content tends to include comparisons, criticism, and decisive opinions.

Finally, for the unsafe sentences, the most commonly used words are the Hebrew words for: 'wants', 'me', 'him', 'now', 'you', 'that I', 'someone' and 'doing'.

These words can indicate bullying, as they seem to suggest that someone is trying to impose his or her will on someone else, and may even imply that this must be done promptly.

We also test the following properties of the audio files against each of the categories.

1. MFCC (Mel Frequency Cepstral Coefficient) is a set of fundamental audio features, using a 20 ms audio frame unit. It gives among other things the noise, speech rate, speech acceleration etc. Spectral contrast - gives the contrast in the audio.

2. Mel-scale spectrogram is a spectrogram in which the frequencies are converted to the mel-scale. The Mel scale is similar to human's ear as they are equal in distance from

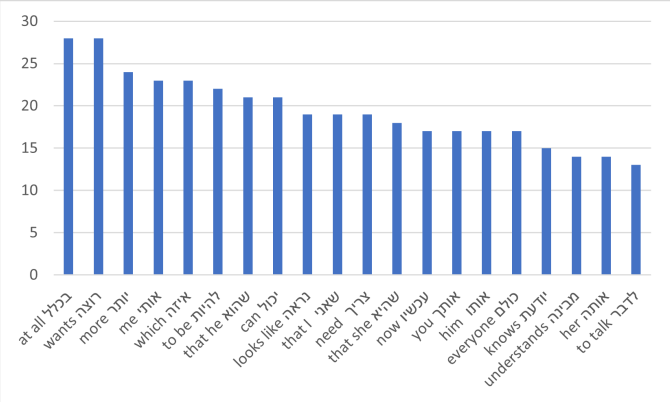

Figure 14: Insulting Sentences Word's Frequency

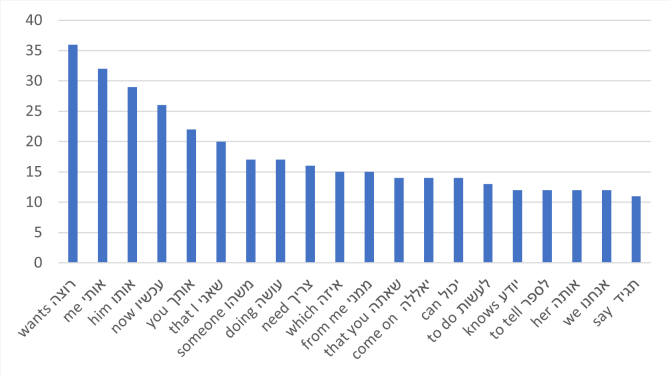

Figure 15: Unsafe Sentences Word's Frequency

one another.

3. Spectral contrast is the difference in amplitude between the spectral peaks and valleys for six subbands for each time frame.

4. Short-time Fourier transform (STFT) is a Fourier transform that takes place around a short time and evaluates the Fourier return on the time-dependent segment. The Chroma feature relates to the twelve different pitch classes, it provides a robust way to describe a similarity measure between audio pieces.

5. Tonnetz is a tonal space representation introduced by Euler. It helps detecting Harmonic Change in Musical Audio - tonation.

Table 9 and Figure 16 present the average value of each characteristic for each category.

As can be seen in the table both Tonnetz and Chroma have the highest value in unsafe speech. This indicates that in unsafe speech there is more variety in the audio and also has more use of tonation. MFCC and Mel has the highest value in neutral speech, probably

Table 9: **Average Value of Each Parameters for Each Category**

Feature	Neutral	Insulting	Unsafe
MFCC	-5.94	-5.54	-5.26
Mel	-22.73	-20.24	-19.84
STFT Chroma	0.568	0.567	0.577
Contrast	19.63	19.75	19.54
Tonnetz	0.0059	0.0038	0.0068

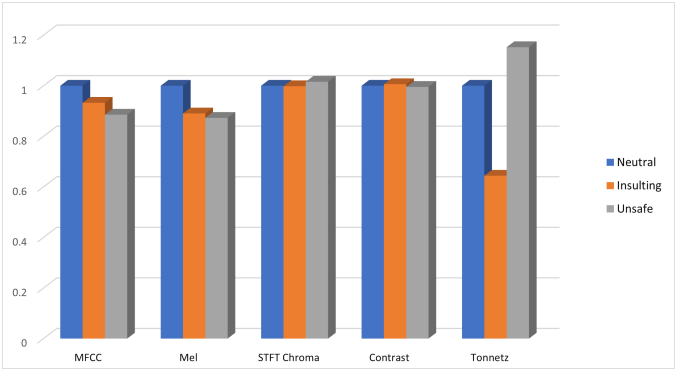

Figure 16: Average Value of Each Audio Parameter, Divided by the Neutral Average Value

because both are related to a normal human's ear. And Contrast has the high values for insulting speech because the contrast in the audio is more noticeable for this kind of sentences.

5.2 Methodology Description for Classification via Hebrew Text and Voice Dataset

After seeing the results of an English text analysis, we wanted to get closer to the real world of our problem, as mentioned above, we want to develop an agent who will help the child with 'special needs' to understand the environment and behave as expected in variety of situations. At this point we also wanted to test whether adding the audio improves our performance. In order to detect insulting and harmful contents given by text and voice contents, we used text recognition methods as well as methods that combine text and audio recognition. The exact hyper parameters used for each method can be found in the project implementation[39].

[39] `https://github.com/MLspecialNeeds/harmful_sentence_detection`

5.3 Methods Used for Classification via Text Features

We started with the database of Hebrew text sentences and applied different methods of machine learning to the sentences, some of which we used in the first part of the work. In addition, we used other diverse methods here, such as BERT, etc., in order to achieve a better level of accuracy. We considered the following classical ML methods for classification via text contents:

1. **MLP** is a fully connected neural network with the following three hidden layers. A general description mentioned above 4.3 We used the first hidden layer size is identical to the input size; it is followed by a hidden layer of size 100 and 50. Each layer is followed by a tanh activation function. It uses a weighted cross-entropy loss function and the ADAM optimizer with a learning rate of 1e-4.

2. **SVM** are a set of supervised learning methods used for classification. Detailed explanation mentioned above in Section 4.3.

3. **KNeighbors** is an unsupervised learning based on a similarity measure of its neighbors. Detailed explanation mentioned above in Section 4.3.

4. **Random Forest** is an ensemble learning method for classification. Detailed explanation mentioned above in Section 4.3

5. **ExtraTrees** combines the predictions from many decision trees.

6. **Logistic Regression** is a model that used when the value of the target variable is categorical in nature. When we had tree categories we used the softmax algorithm.

7. **NB** is a classification technique assuming that the predictors are independent based on Bayes' Theorem. Detailed explained above in Section 4.3

8. **Voting** is a combination of all above machine learning classifiers and uses a majority vote or the average predicted probabilities to predict the labels. Details explained above in Section 4.3

9. **Bert** Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique developed for natural language processing (NLP) pre-training. The implementation of transformer's two-way training, a popular attention model for language models, is BERT's main technical innovation. Transformer, is an attentional mechanism that learns the relationships between words (or subwords) in a text. It has two separate mechanisms - an encoder that reads the text input and a decoder that produces a prediction for the task.

To improve the classification accuracy, we used pre-trained embedding models for the text inputs. Since our dataset is composed of Hebrew sentences, we use the pre-trained HeBERT model [29], in order to transform the sentences the embedded structure. HebBERT is based on the BERT architecture [78]. Like BERT it is used for diverse NLP tasks, and specially for sentiment analysis. As described in Section 5.6, using the classical ML methods on the embedded inputs improved the accuracy of the classification.

5.4 Methods Used for Classification via Audio

We proceed by describing the details of the classification methods performed by using audio features only. In particular, we used two released Wav2Vec 2.0 pretrained models to transform the audio inputs into a vector embedding structure. We first review the two Wav2Vec models, then we introduce how we use these models and compare them to other baseline methods.

5.4.1 FSFM Classifier

The FSFM model is a Multi-Layer Perceptron (MLP) network from [205], using five-audio features represented as a vector of length 193. The model consists of 4 fully connected layers with ReLU activation function and dropout after each layer, followed by a final layer's uses a softmax activation function. The model uses a weighted categorical cross entropy loss function and the ADAM optimizer.

5.4.2 RNN Classifier

A Recurrent Neural Network (RNN) model. This model is trained on Mel-Frequency Cepstral Coefficients (MFCC), using 20ms 15 audio frames units [193] obtained from the audio samples. The model consists of an LSTM cell with a ReLU activation function, which is followed by a dropout of 0.3, and a fully-connected layer with a sigmoid activation function (or softmax for three classes). It uses a weighted cross-entropy loss function and the ADAM optimizer with a learning rate of 0.0001.

5.4.3 Wav2Vec 2.0 Pretrained models

The Wav2Vec 2.0 model introduced by Baevski et al. [17], is a framework for self-supervised learning of vector representation from speech audio by pretraining on large quantities of audio data developed by Facebook AI. The model attempts to recover a randomly masked portion of the encoded audio feature. The model consists of three main modules. The first module is a feature encoder; it is composed of a 1-d Convolutional neural network encoder, which down-samples the input raw waveform $\mathcal{X}$ to latent speech representation of 25ms each $\mathcal{Z}$ in T time

steps. The second module is a contextualized encoder, which consists of several transformer encoder blocks, transforms the latent representations $\mathcal{Z}$ into contextualized representations $\mathcal{C}$. In addition, there is the quantization module, which takes the speech representation $\mathcal{Z}$ and discretizes them into a finite set of quantized representations $\mathcal{Q}$ by matching them with a codebook for selecting the most appropriate representation of the audio. The objective is to identify these quantized representations of the masked features using the output of the contextualized network $\mathcal{C}$ for each masked time step T by using the contrastive loss function.

After its pretraining on unlabeled audio data, the model can be fine-tuned on labeled data to be used for downstream tasks.

In this work, we compared the accuracy of insulting and unsafe detection using text and audio contents of spoken sentences in different situations. In particular, for the detection through audio, we used both the Wav2Vec 2.0 Base model, called *Wav2vec Base*, and a model that was fine-tuned on Speech Emotion Recognition (SER) task, called *Wav2Vec for Emotions*. The vector embedding sizes are 768 and 1024, respectively. In addition, we fine-tune both models using our data-set.

Wav2Vec 2.0 Base The Wav2Vec 2.0 Base model [40], is pretrained on the Librispeech dataset without fine-tuning.

Wav2vec 2.0 Emotion A pretrained model Ehcalabres[41]. The basic model is a Wav2Vec 2.0 xlsr-53 model [42] that was fine-tuned on English using the Common audio data-set [244] which currently consists of 7,335 hours in 60 languages of transcribed speech. This model was fine-tuned on the RAVDESS data-set [229], a multi-modal database of emotional speech and song which contains 1440 hours of samples in eight different emotions classes, recording professional actors, in English.

Wav2Vec 2.0 Fine-Tuning During the fine-tuning process, we take the context representation of our data from the pre-trained models, starting with an average pooling layer that calculates an averaged vector according to the time dimension, add is followed by a fully connected layer with Tanh activation function. Finally, there is a fully connected layer for the classification task. Since the Wav2Vec 2.0 model was used as a feature extractor, the weights of the features encoder module of the pre-trained model were not changed during the fine-tuning process. This fine-tuning architecture is inspired by [218] due to its similarity to

[40] `https://huggingface.co/facebook/wav2vec2-base`

[41] `https://huggingface.co/ehcalabres/wav2vec2-lg-xlsr-en-speech-emotion-recognition`

[42] `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`

our task and for achieving satisfactory results on their tasks. The hyper-parameters used for fine-tuning are represented in Table 10.

Table 10: Hyper-parameters used for fine-tuning

Parameters	**Base Model**	**Emotion Model**
Sample frequency	16k Hz	16k Hz
Learning rate	2e-5	1e-4
Training epochs	10	10
Training batch size	3	3
Gradient accumulation steps	2	2
Total train batch size	6	6

Finally, we consider a model that uses the logits of the 8 emotions classes extracted from the emotion model (without fine-tuning on our data). This model is referred to as *Wav2vec Emotion Vector* in our simulation results description.

For training all these extracted features from the Wav2vec models we use the following MLP model, which combines text and audio features, in order to check the effect of analysis by both voice and text.

5.5 Combined Text and Audio Methods

In the next set of experiments, we use both audio and text features. We assume that combining audio and text together will enhance the result of the classification task.

For training both features we build a *Wav2vec Emotion Vector* model, that shrinks the given audio wav2vec embedded vector into a size of 100, and then we concatenate it with the text hebBERT embedded vector of size 768, which is followed by the text MLP model. The described model can be seen in Figure 17.

The details of the combined model, as well as the motivation for this model, are described as follows. The aim of the combined model is to utilize both textual and audio information in order to achieve greater accuracy in detecting challenging situations. We first process the text data using vectorization by Bert and a fully connected layer, with 100 neurons and tanh activation function. In addition, the audio was processed using wav2vec and then passed through two fully connected layers with tanh activation function, and then a 30% dropout layer was added after each FC layer. Then, both outputs (the textual-based output and the audio-based output) were merged, and 3 additional fully connected layers were used, followed by a softmax layer for the dangerous/insulting situation detection. This architecture was

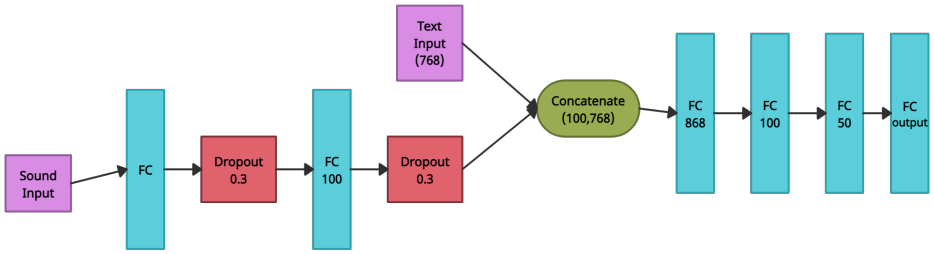

Figure 17: An illustration of the text & audio model.

found to give the most accurate results, as described in the next section. For additional information, we refer the reader to our github site[43]

Based on these various machine learning methods, in the next section we compare the accuracy of the different models: text-based models, audio-based models, and the combined text-and-audio based model, on preset challenging situations.

5.6 Experimental Results

In the following section, we describe our experimental results for detecting insulting and unsafe sentence contents, using text, audio, or both sources, and applying the machine learning methods described above. In all our experiments, we test all models using 5 fold cross validation on our collected data-set, and the accuracy presented in the results table is the accuracy on the test set, in all our experiments. In addition, for all the experiments using DNN, we run a model with 100 epochs. It should be noted that in the study described in Section 4.4, based on text-based learning in classical ML methods, we used a test method in which we ran 50 runs and we divided each lecture into 90% in the training group and 10% in the test group, while in the study described in this section, we used cross validation. The reason for this difference is that in the study described in this section, the training process based on deep networks took longer. Therefore we preferred the cross validation method, with a division of 80% in training and 20% in test, which enables a faster testing process. First, we utilize classical machine learning methods for insulting and unsafe sentence detection, using the text sources. We used weighted cross-entropy loss function in all methods except the KNN and voting, our results are presented in Table 11 for insulting sentences detection, in Table 12 for unsafe situation detection, and in Table 13 for all three categories classification.

In Table 11 we can see that when detecting insulting sentences, we got the best result when we used Bert with the MLP algorithm.

In Table 12 we can see that when detecting unsafe sentences, we got the best result when

[43] `https://github.com/MLspecialNeeds/harmful_sentence_detection`

Table 11: Accuracy of insulting sentences detection based on text only

Model	Tfidf	BERT
MLP	66.54	**80.66**
SVM	71.91	79.05
KNeighbors	65.46	74.48
RandomForest	70.19	78.35
ExtraTrees	70.19	76.53
Logistic Regression	71.32	76.36
NB	70.19	-
Voting (without MLP)	71.69	78.40

Table 12: Accuracy of unsafe sentences detection based on text only.

Model	Tfidf	BERT
MLP	67.46	**78.33**
SVM	71.26	76.28
KNeighbors	69.04	71.61
RandomForest	68.51	74.12
ExtraTrees	69.39	74.12
Logistic Regression	70.56	73.42
NB	70.21	-
Voting (without MLP)	71.26	75.29

we used Bert with MLP algorithm. $0, 2, 4$

In Table 13 we can see that when all 3 categories were combined, we got the best result when we used Bert with SVM algorithm.

We proceed by examining the ability of insulting and unsafe sentences detection by using the speech audio files, and by combining text and audio data. It was assumed that adding audio with text would enhance the results of each of them separately, but not all the additions produced the expected results.

Tables 14, 15 and 16 describe the experiment's results for various combinations of text and wave embedded vectors. The methods are compared to a baseline method which uses a randomly generated vector of length 768 sampled from the same distribution as the wav2vec vectors (we will call this *Random Vector*), that was used instead of the wav2vec embedding vectors.

Table 13: Accuracy, Precision and Recall of classifiers for all three categories based on text only.

Model	**Tfidf**				**BERT**			
	ACC	**F1**	**Prec**	**Rec**	**ACC**	**F1**	**Prec**	**Rec**
MLP	53.08	52.8	53.06	53.1	61.98	61.72	62.24	61.68
SVM	57.27	56.96	57.12	57.04	**64.59**	64.08	64.23	64.31
KNeighbors	51.66	51.03	53.01	51.56	57.04	55.96	56.47	56.43
Random Forest	53.94	52.52	54.12	53.29	63.77	62.46	63.62	62.96
ExtraTrees	55.73	54.87	56.25	55.43	60.97	59.7	60.59	60.13
Softmax Regression	56.48	58.28	56.29	56.38	63.58	63.1	63.2	63.3
NB	56.85	-						
Voting (without MLP)	57.23	56.28	57.95 56.77		63.7	62.4	63.6	62.9

Table 14: Accuracy of insulting speech detection.

Model	**fine-tune on SER data**	**fine-tune on our data**	**Audio only**	**Audio + BERT**
FSFM [205]	N/A	N/A	58.38	79.7
RNN	N/A	N/A	58.65	78.36
Wav2vec Base	✗	✗	58.65	80.67
Wav2vec For Emotions	✓	✗	61.22	80.45
Wav2vec Base	✗	✓	64.66	80.34
Wav2vec For Emotions	✓	✓	**66.06**	77.66
Wav2vec Emotion Vector	✓	✗	50.7	**80.88**
Random Vector	N/A	N/A	48.66	77.01
BERT alone	N/A	N/A	—	80.66

We hypothesised that adding the audio features to the text data would improve the accuracy of the results. Indeed, when combining audio to text, some improvement was reached:

Table 15: Accuracy of unsafe speech detection.

Model	fine-tune on SER data	fine-tune on our data	Audio only	Audio + BERT
FSFM [205]	N/A	N/A	61.57	78.33
RNN	N/A	N/A	59.81	75.39
Wav2vec Base	✗	✗	63.32	78.79
Wav2vec For Emotions	✓	✗	61.79	**80.02**
Wav2vec Base	✗	✓	**68.98**	74.88
Wav2vec For Emotions	✓	✓	67.87	77.68
Wav2vec Emotion Vector	✓	✗	49.36	78.68
Random Vector	N/A	N/A	51.05	74.53
BERT alone	N/A	N/A	—	78.33

Table 16: Accuracy of classifiers on all three categories based on speech.

Model	fine-tune on SER data	fine-tune on our data	Audio only	Audio + BERT
FSFM [205]	N/A	N/A	42.84	64.22
RNN	N/A	N/A	39.55	60.07
Wav2vec Base	✗	✗	45.9	**65.79**
Wav2vec For Emotions	✓	✗	44.41	65.15
Wav2vec Base	✗	✓	47.1	64.78
Wav2vec For Emotions	✓	✓	**49.05**	61.42
Wav2vec Emotion Vector	✓	✗	33.6	63.25
Random Vector	N/A	N/A	35.17	60.52
BERT alone	N/A	N/A	—	63.43

approximately two percents of accuracy for each of the binary classification, and about 0.22 percent for the three categories classification. We believe that we only see a relatively small

improvement since most of the information can be obtained from text. Indeed, the results of text-only based learning were much higher than the results of audio-only based learning. Another explanation for this phenomenon lies in the fact that there are unsafe situations in which the offender will have an interest in speaking in a normal tone in order to hide the danger. Similarly, insulting speech has no unique sound characteristics that set it apart from ordinary conversation.

The random vector sometimes provided the best results because not only did the voice of the audio data not improve the result, it was worsened. When the audio data was combined with BERT we saw a benefit. However, the emotion vector didn't help us thjo recognize risky sentences.

Next, we would like to compare the different embedding variations used as inputs of the DNN. The results presented in Tables 14, 15 and 16 clearly show that fine-tuning of our data helps when using audio alone, but when we combine it with text, it decreases the accuracy. On the other hand, fine-tuning of the embedding audio vectors on RAVDESS emotional dataset [229] did improve the accuracy of the learning process. We believe that using fine-tuning on our own datasets as an early stage on audio only, causes the model to become too adapted to our dataset, negatively affecting the final training with relevant text. Training the model without this stage, and when using both audio and text, achieved the best result.

6 Research Contributions

The current research has significant contributions in two different research areas. First, it can help in the development of automated agent to assist children with special needs. Second, it may advance the research of text classification and threat detection using text and voice signals. We would like to expand both of these aspects.

A child with special needs has difficulty understanding the social nuances of his environment. In addition, because of the vulnerability of these children their protection becomes more important. Parents of these children want to afford them independence, like "normal" children, but, on the other hand, they know that dangers lurk outside. There are cases where people exploit the misunderstanding of these children and hurt them. From our discourse with parents of these vulnerable children, there is a great need for solutions that can allow them to be less concerned about the lurking dangers.

The agent we intend to develop can help in various ways in daily life. In order to help protect children with special needs, the agent can assist them in understanding whether or not a special situation is a risk of danger or not.

The second important issue for such children is improvement in social skills. This area is very wide and includes both active and passive behavior. If a child says something that could hurt another, the agent will recognize the situation and offer the child different ways to correct what has been said or rectify the situation. When a child says:"you are stupid" it will probably hurt person to whom he said it. Children can be tactless at best and cruel at worst. Therefore, the child can be exposed to a variety of social utterances' where, for example: "what a beautiful shirt" can be said as a compliment or as a sarcastic remark. Towards this goal, the assisting agent will understand the complicated statement and signal the child, with the information he lacks to correctly understand the situation.

In respect to Machine Learning research, our contribution is in developing tools to determine special conditions given text and voice signals.

In our ongoing research, we developed new datasets that are constructed from a large pool of sentences which can be utilized to allow similar projects on sentiment analysis. We describe our datasets in Section 4.1. The datasets will be made available upon request.

The experiments presented in the previous section reveal the tremendous potential of the concept of using deep learning to identify bullying situations or emotions or sarcasm. We intend to improve the existing algorithm in order to obtain the best results for the benefit of these children.

The current work is different from studies on classical sentiment analysis. This is because in sentiment analysis there is emotion detection concerning the writer, whereas in our work,

focus is on detecting the sentences that cause the listener to get insulted or bulled. It is also different from hate-speech detection, because the insulting sentences in the domain of this work can be the result of innocent intentions, where most cases do not contain sentences that are considered hate speech.

7 Conclusions

Until now, we handled the challenge of detecting insulting and unsafe situations using text and audio speech contents and using vector embedding for both text and audio. This challenge can be viewed as a classification problem into the following three classes: neutral sentences, sentences consisting of insulting contents, and sentences indicating unsafe situations. We concentrated on situations relevant for children, and in particular, for children with special needs.

In our experiments, we found that adding the embedded wave information to the text information, only slightly improved the overall accuracy. This may be since in unsafe situations the main information conveyed is related to the words being spoken and not as much to the way they are spoken or other characteristics of the audio. This may be especially true when it comes to a dialogue between people who are not close as family members. In addition, we found that fine-tuning of the embedded wave vectors by our dataset reduces the final accuracy. Also, the rise of CAs and their applications can have significant influence on our future life. Some of these applications are positive and even crucial, such as health support or social support; others can be beneficial to business and companies; and others should be monitored or even avoided for moral reasons. The limits of fair use of CAs and the technological tools to enforce these limits should be discussed and developed in future research.

We proceeded by describing additional aspects of the work that we intend to consider in future research, towards the goal of developing an automated agent that will be able to assist children with special needs. First of all, in this study, we collected text and audio datasets concentrating on unsafe and insulting situations related to children, and especially related to special needs children. It is interesting to check whether the same results will be obtained for other unsafe situations, such as violence situations between adults or in detecting domestic violence. Another interesting open question is to check what the audio is effect in different cultures and languages. In addition, it is interesting to examine whether adding video films and/or pictures of the examined events will increase the risk detection accuracy.

Moving from a simple sentence to a brief conversation raises the complexity of the analysis required, since several sentences may have different meanings for different conversation contents. Moreover, it is important that the assisting agent will be able to recommend to the

child the response that should be taken under the current circumstances.

In this study, we show how to detect challenging (bullying, abusing, and insulting) situations, using the content of text and audio sentences. As a follow-up to this study, we propose to test RNN-based methods for identifying a challenging situations by examining the whole flow of the conversation, for which data from full and long conversations can be used.

Another direction for future work is development of a supporting agent that will advise the child how to act in various challenging situations. Towards this goal, we suggest that proper conversational datasets, probably conversations taken from movies, can be used. In addition, human subjects can be asked for the appropriate response that should be made in the current circumstances. Possible reactions to be considered could include: ignore, answer, walk away, call for help, etc. In order to build the human-based responses, crowdsourcing can be used to suggest appropriate responses given several challenging situations, and/or to score given responses. Based on the gathered data, a machine learning model (probably an RL-based model) can be used to recommend actions to a child, or to suggest contacting the child's parents or the emergency services in cases in which a dangerous situation is detected.

A Text Based Methods - Confusion Matrices

In this section, we provide the confusion matrices of the text-based model examined in Section 4.4.

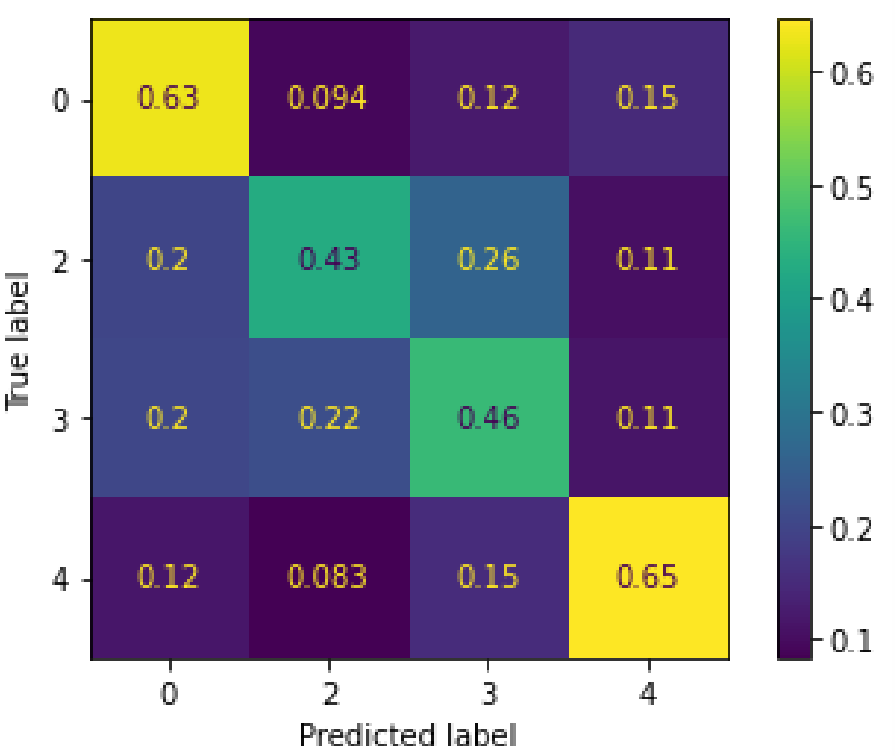

Figure 18: Ridge Classifier Confusion Matrix

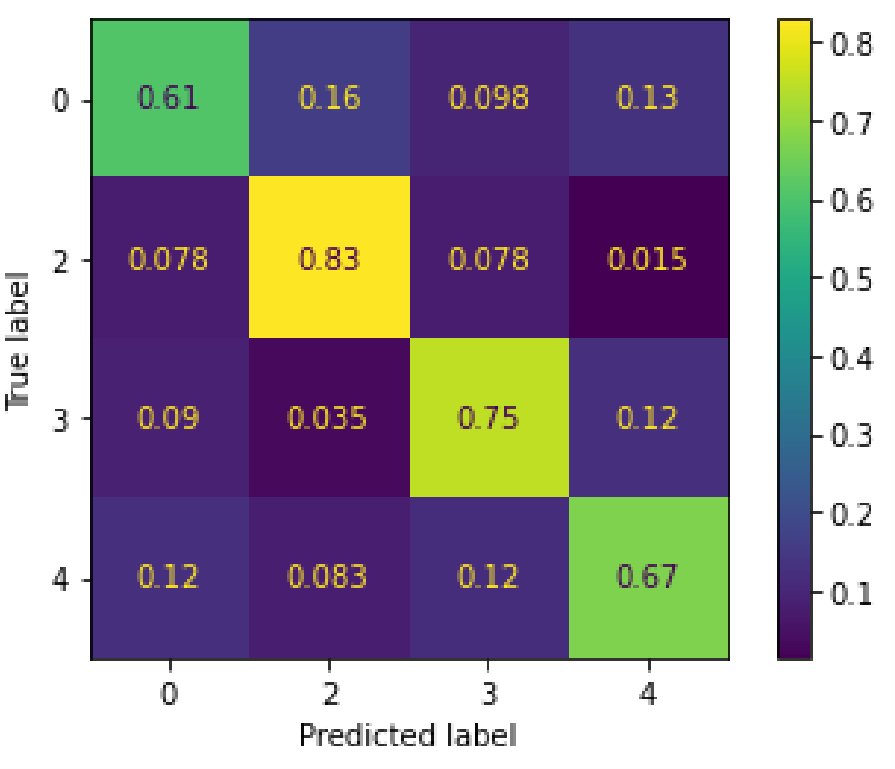

Figure 19: SVM Classifier Confusion Matrix

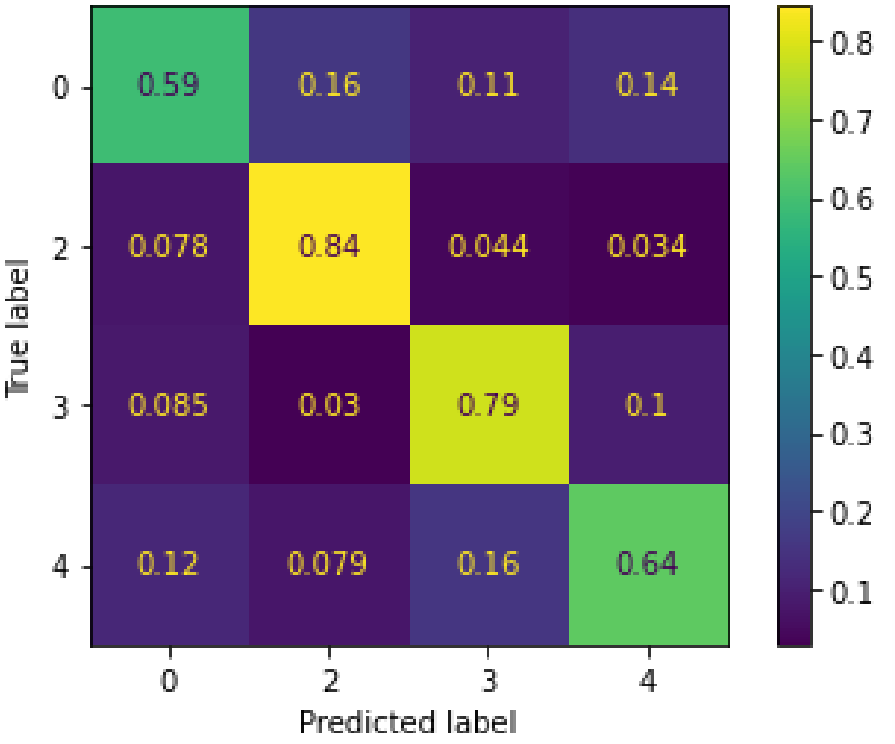

Figure 20: KNN Classifier Confusion Matrix

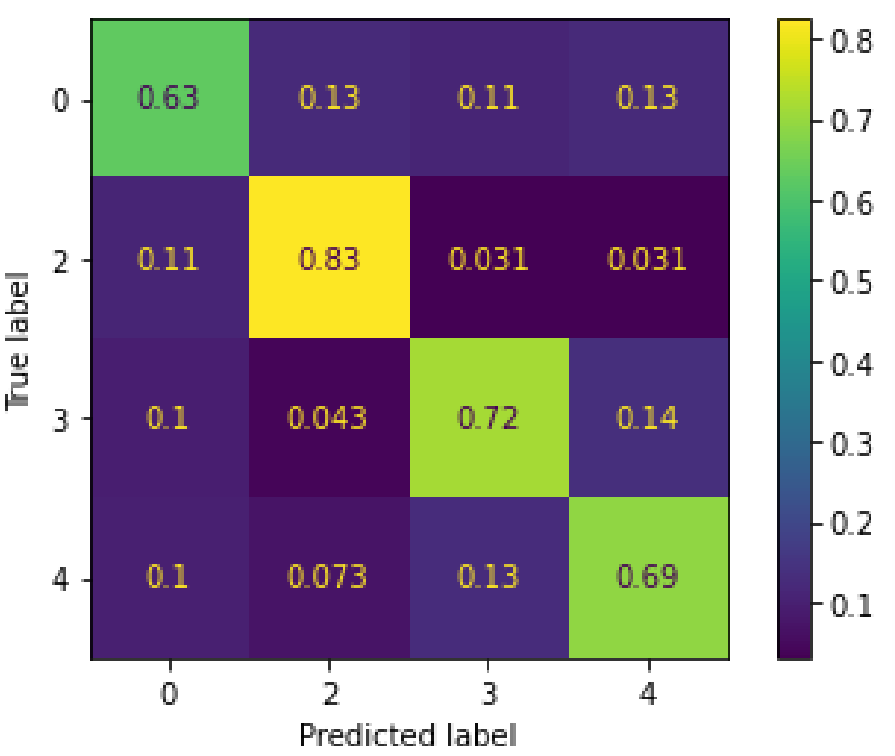

Figure 21: Extra Trees Classifier Confusion Matrix

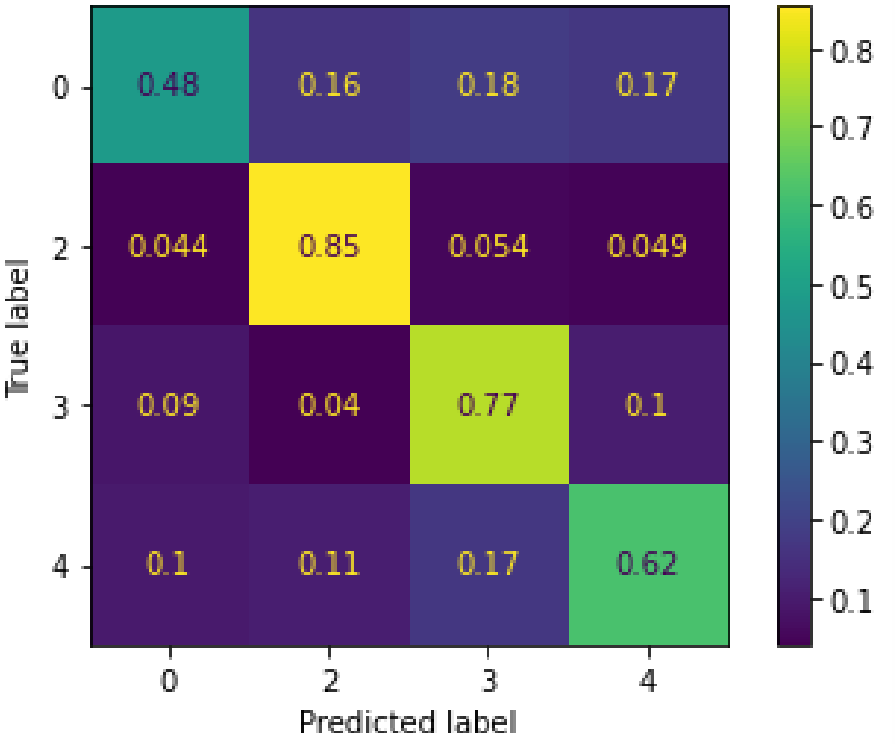

Figure 22: Bayes Classifier Confusion Matrix

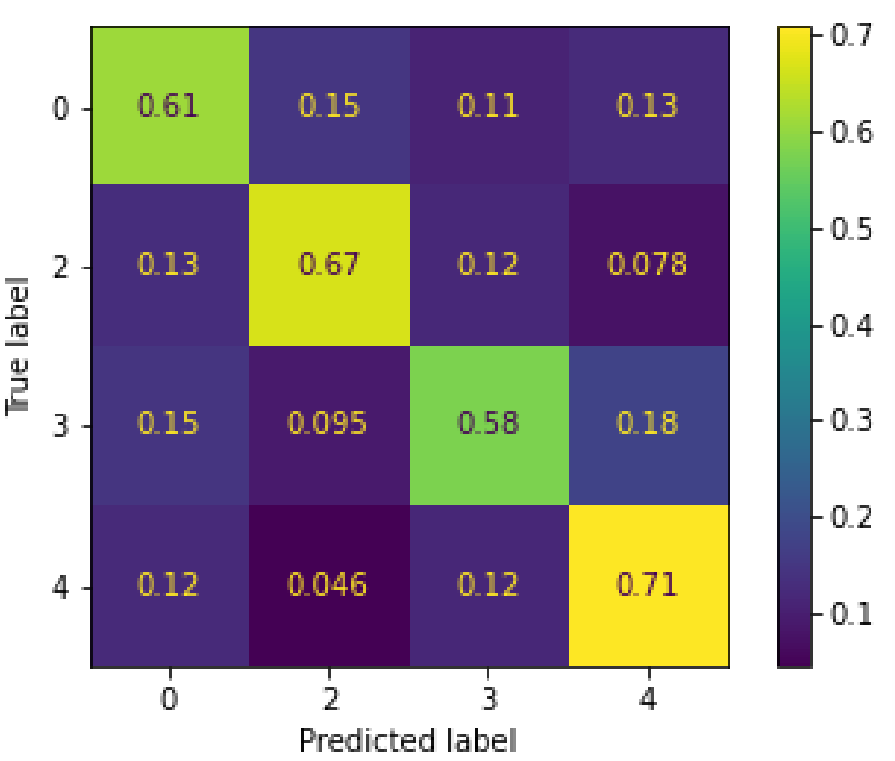

Figure 23: Voting Classifier Confusion Matrix

References

[1] Sarder M. A. Ecactive embodied conversational agent for mental health intervention. Master's thesis, Delft University of Technology, 2018.

[2] H. Chen A. Abbasi and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26, 2008.

[3] Ahmad Abdellatif, Khaled Badran, and Emad Shihab. Msrbot: Using bots to answer questions from software repositories. *Empirical Software Engineering*, 25(3):1834–1863, 2020.

[4] Sameera A. Abdul-Kader and John Woods. Survey on chatbot design techniques in speech conversation systems. *(IJACSA) International Journal of Advanced Computer Science and Applications*, 6, 2015.

[5] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

[6] Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.

[7] Eleni Adamopoulou and Lefteris Moussiades. An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 373–383. Springer, 2020.

[8] Terrence Adams. Ai-powered social bots. arXiv preprint https://arxiv.org/abs/1706.05143.

[9] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

[10] Amit Agarwal. How to write a twitter bot in 5 minutes. https://www.labnol.org/internet/write-twitter-bot/27902/.

[11] Francesco Agostaro, Agnese Augello, Giovanni Pilato, Giorgio Vassallo, and Salvatore Gaglio. A conversational agent based on a conceptual interpretation of a data driven semantic space. In *Congress of the Italian Association for Artificial Intelligence*, pages 381–392. Springer, 2005.

[12] Cristina Catalán Aguirre, Carlos Delgado Kloos, Carlos Alario-Hoyos, and Pedro J Muñoz-Merino. Supporting a mooc through a conversational agent. design of a first prototype. In *2018 International Symposium on Computers in Education (SIIE)*, pages 1–6. IEEE, 2018.

[13] Nahdatul Akma Ahmad, Mohamad Hafiz Che Hamid, Azaliza Zainal, Muhammad Fairuz Abd Rauf, and Zuraidy Adnan. Review of chatbots design techniques. *International Journal of Computer Applicationss*, 181:56–67, 2018.

[14] Addi Ait-Mlouk and Lili Jiang. Kbot: a knowledge graph based chatbot for natural language understanding over linked data. *IEEE Access*, 8:149220–149230, 2020.

[15] Icek Ajzen. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211, 1991.

[16] M. Akhtar, A. Kumar, D. Ghosal, A.Ekbal, and P. Bhattacharyya. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *In Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2017.

[17] Baevski Alexei, Zhou Henry, Mohamed Abdelrahman, and Auli Michael. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[18] Kourosh Alizadeh. Limitations of twitter data issues to be aware of when using twitter text data. https://towardsdatascience.com/limitations-of-twitter-data-94954850cacf.

[19] Leslie AM and Frith U. Autistic children's understanding of seeing, knowing, and believing. *Brit J Dev Psychol.*, 6:315–24, 1988.

[20] Amazon. Amazon mechanical turk. https://www.mturk.com, Accessed on 13.03.2022.

[21] David Ameixa, Luisa Coheur, and Rua Alves Redol. From subtitles to human interactions: introducing the subtle corpus. Technical report, Technical report, 2013.

[22] Azaria Amos and Nivasch Keren. Saif: A correction-detection deep-learning architecture for personal assistants. *Sensors*, 20(19):5577, 2020.

[23] and lati B, Boccanfuso L, Huang CM, Mademtzi M, Qin M, and et al. Improving social skills in children with asd using a long-term in-home social robot. *Sci. Robot.*, 3:eaat 7544, 2018.

[24] VA Arlington. *American Psychiatric Association . Diagnostic and statistical manual of mental disorders (DSM-V)*. American Psychiatric Publishing, 2013.

[25] Lokman A.S. and Ameedeen M.A. Modern chatbot systems: A technical review. In *Proceedings of the Future Technologies Conference (FTC)*, volume 881, pages 1012–1023. Springer, 2019.

[26] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. Affective neural response generation. In *European Conference on Information Retrieval*, pages 154–166. Springer, 2018.

[27] Dennis Assenmacher, Lena Clever, and Lena Frischlichy. Demystifying social bots: On the intelligence of automated social media actors. *Social Media + Society*, pages 1–14, 2020.

[28] Google Assistant. Google assistant, your own personal google. https://assistant.google.com/.

[29] Chriqui Avihay and Yahav Inbal. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*, 2021.

[30] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features and classification schemes and databases. *Pattern Recog.*, 44:572–587, 2011.

[31] Amos Azaria, Jayant Krishnamurthy, and Tom Mitchell. Instructable intelligent personal agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[32] Amos Azaria, Ariella Richardson, and Sarit Kraus. An agent for deception detection in discussion based environments. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 218–227. ACM, 2015.

[33] Amos Azaria, Shashank Srivastava, Jayant Krishnamurthy, Igor Labutov, and Tom M Mitchell. An agent for learning new natural language commands. *Autonomous Agents and Multi-Agent Systems*, 34(1):1–27, 2020.

[34] Roger Azevedo, Ronald S Landis, Reza Feyzi-Behnagh, Melissa Duffy, Gregory Trevors, Jason M Harley, François Bouchet, Jonathan Burlison, Michelle Taub, Nicole Pacampara, et al. The effectiveness of pedagogical agents' prompting and feedback in facilitating co-adapted learning with metatutor. In *International conference on intelligent tutoring systems*, pages 212–221. Springer, 2012.

[35] Huskens B., Palmen A., der Werff M. V., Lourens T., and Barakova E. Improving collaborative play between children with autism spectrum disorders and their siblings: The effectiveness of a robot-mediated intervention based on lego therapy. *Journal of Autism and Developmental*, 2014.

[36] Vanderborght B., Simut, R., Saldien, J., Saldien, J., Pop, C., Rusu, A.S., Pineta, S., Lefeber, D., David, and D.O. Using the social robot probo as a social story telling agent for children with asd. *Interact. Stud.*, 13:348–372, 2012.

[37] Rafael E Banchs and Haizhou Li. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42, 2012.

[38] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72, 2005.

[39] Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. Plato: Pre-trained dialogue generation model with discrete latent variable. In *ACL*, page 85–96, 2020.

[40] S. Baron-Cohen, A. M.Leslie, and U. Frith. Does the autistic child have a theory of mind? *Cognition*, 21:37–46, 1985.

[41] Rodrigo Bavaresco, Diórgenes Silveira, Eduardo Reis, Jorge Barbosa, Rodrigo Righi, Cristiano Costa, Rodolfo Antunes, Marcio Gomes, Clauter Gatti, Mariangela Vanzin, et al. Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review*, 36:100239, 2020.

[42] Aryel Beck, Brett Stevens, Kim A Bard, and Lola Cañamero. Emotional body language displayed by artificial agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(1):1–29, 2012.

[43] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786, 2007.

[44] D. Bertero, F. B. Siddique, C. S. Wu, Y. Wan, R. H. Chan, and P. Fung. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *In*

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), 2016.

[45] Heena Reyaz Bhat, Tanveer Ahmad Lone, and Zubair M Paul. Cortana-intelligent personal digital assistant: a review. *International Journal of Advanced Research in Computer Science*, 8(7):55–57, 2017.

[46] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[47] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walquer H Hill, David R Krathwohl, et al. Taxonomy of educational objetives: the classification of educational goals: handbook i: cognitive domain. Technical report, New York, US: D. Mckay, 1956.

[48] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.

[49] Bhriguraj Borah, Dhrubajyoti Pathak, Priyankoo Sarmah, Bidisha Som, and Sukumar Nandi. Survey of textbased chatbot in perspective of recent technologies. In *International Conference on Computational Intelligence, Communications, and Business Analytics*, pages 84–96. Springer, 2018.

[50] Bianca Bosker. Siri rising: The inside story of siri's origins–and why she could overshadow the iphone. *Huffington Post*, 2013.

[51] Sofiane Boucenna, Antonio Narzisi, Elodie Tilmont, Filippo Muratori, Giovanni Pioggia, David Cohen, and Mohamed Chetouani. Interactive technologies for autistic children: A review. *Cognitive Computation*, 6:722–740, 2014.

[52] Sofiane Boucenna, Antonio Narzisi, Elodie Tilmont, Filippo Muratori, Giovanni Pioggia, David Cohen, and Mohamed Chetouani. Interactive technologies for autistic children: A review. *Cognitive Computation*, 6(4):722–740, 2014.

[53] Cynthia Breazeal. Social robots: from research to commercialization. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1–1, 2017.

[54] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *EMNLP*, 2018.

[55] J.K. Burgoon, L.K. Guerrero, and V. Manusov. *Nonverbal signals*, pages 239–282. SAGE Publications, Thousand Oaks, CA, USA, 2011.

[56] Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Daniel Duckworth, Semih Yavuz, Ben Goodrich, Amit Dubey, Andy Cedilnik, and Kyu-Young Kim. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *EMNLP-IJCNLP*, 2019.

[57] Grossarda C., Grynspanb O., Serretc S.and Jouenb AL., Baillyb K., and Cohen D. Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (asd). *Computers & Education*, 113:195–211, 2017.

[58] Hughes C. and Leekam S. What are the links between theory of mind and social relations? review, reflections and new directions for studies of typical and atypical development. *Soc. Dev.*, 13:590–619, 2004.

[59] Maiano C., Normand CL.and Salvas MC.and Moullec G., and Aime A. Prevalence of school bullying among youth with autism spectrum disorders: A systematic review and meta-analysis. *Autism Research*, 9:601–615, 2016.

[60] William Cai, Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph Jay Williams, and Sharad Goel. Mathbot: A personalized conversational agent for learning math. *Published to ACM*, 2019.

[61] Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Joty Shafiq Rayhan, Yin Leng Theng, and Rifat Atun. Conversational agents in health care: Scoping review and conceptual analysis. *Journal of medical Internet research*, 22(8):e17158, 2020.

[62] Valentina Carfora, Francesca Di Massimo, Rebecca Rastelli, Patrizia Catellani, and Marco Piastra. Dialogue management in conversational agents through psychology of persuasion and machine learning. *Multimedia Tools and Applications volume*, 79:35949–35971, 2020.

[63] Yixuan Chai, Guohua Liu, Ziwei Jin, and Donghong Sun. How to keep an online learning chatbot from being corrupted. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[64] Ana Paula Chaves and Marco Aurelio Gerosa. How should my chatbot interact? a survey on human-chatbot interaction design, 2020. arXiv preprint https://arxiv.org/abs/1904.02743.

[65] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017.

[66] Jinyin Chen, Yangyang Wu, Chengyu Jia, Haibin Zheng, and Guohan Huang. Customizable text generation via conditional text generative adversarial network. *Neurocomputing*, 416:125–135, 2020.

[67] Merav Chkroun and Amos Azaria. Lia: A virtual assistant that can be taught new commands by speech. *International Journal of Human–Computer Interaction*, pages 1–12, 2019.

[68] Leigh Michael Harry Clark, Nadia Pantidi, Orla Cooney, Philip R Doyle Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, and Cosmin Munteanu. What makes a good conversation?: Challenges in designing truly conversational agents. In *In Proceedings of the 2019 CHI Conference*, 2019.

[69] Kenneth Mark Colby. Ten criticisms of parry. *ACM SIGART Bulletin*, 48:5–9, 1974.

[70] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[71] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, System Demonstrations*, pages 97–102, 2017.

[72] Maral Dadvar, Rudolf Berend Trieschnigg, and Franciska MG de Jong. Expert knowledge for automatic detection of bullies in social networks. In *25th Benelux Conference on Artificial Intelligence, BNAIC 2013*, pages 57–64. TU Delft, 2013.

[73] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *arXiv preprint arXiv:1106.3077*, 2011.

[74] daniel Peterschmidt. How to make a twitter bot in under an hour even if you don't code that often. https://medium.com/science-friday-footnotes/how-to-make-a-twitter-bot-in-under-an-hour-259597558acf.

[75] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2951–2960, 2017.

[76] J. Deng, Z. Zhang, F. Eyben, and B. Schuller. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Process*, 21:1068–1072, 2014.

[77] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54, 2021.

[78] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[79] Sanjay Dhanda. How chatbots will transform the retail industry. *Juniper Research*, 2018.

[80] Stephan Diederich, Alfred Benedikt Brendel, and Lutz M. Kolbe. On conversational agents in information systems research: Analyzing the past to guide future work. In *14th International Conference on Wirtschaftsinformatiks*, 2019.

[81] Stephan Diederich, Alfred Benedikt Brendel, and Lutz M Kolbe. Towards a taxonomy of platforms for conversational agent design. In *WI 2019*, 2019.

[82] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems. In *ICLR*, 2016.

[83] Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech and Language*, 59:123–156, 2020.

[84] Begoli E. Procedural reasoning system (prs) architecture for agent-mediated behavioral interventions. In *Research Gate*, 2014.

[85] Roger A Edwards, Timothy Bickmore, Lucia Jenkins, Mary Foley, and Justin Manjourides. Use of an interactive computer agent to support breastfeeding. *Maternal and child health journal*, 17(10):1961–1968, 2013.

[86] Costa A.P. et al. More attention and less repetitive and stereotyped behaviors using a robot with children with autism. In *Proc. 27th IEEE Int. Symp. Robot Human Interactive Commun.*, 2018.

[87] Noroozi F., Sapiński T., Kamińska D., and Anbarjafari G. Vocal-based emotion recognition using random forests and decision tree. *Int. J. Speech Technol.*, 25:1–8, 2017.

[88] Ralf Fabian and Marcu Alexandru-Nicolae. Natural language processing implementation on romanian chatbot. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*. WSEAS, 2009.

[89] Ahmed Fadhil, Yunlong Wang, and Harald Reiterer. Assistive conversational agent for health coaching: a validation study. *Methods of information in medicine*, 58(1):009–023, 2019.

[90] Wang Yi Fei and Stephen Petrina. Using learning analytics to understand the design of an intelligent language tutor–chatbot lucy. *Editorial Preface*, 4:124–131, 2013.

[91] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 132:138–161, 2019. https://www.sciencedirect.com/science/article/pii/S1071581918305238.

[92] Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. An intelligent discussion-bot for answering student queries in threaded discussions. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 171–177, 2006.

[93] Libby Ferland and Wilma Koutstaal. How's your day look? the (un)expected sociolinguistic effects of user modeling in a conversational agent. In *CHI 2020)*, pages 482–489, 2020.

[94] A. Fernandes. Nlp, nlu, nlg and how chatbots work. https://chatbotslife.com/nlp-nlu-nlg-and-how-chatbots-work-dd7861dfc9df.

[95] Emilio Ferrara, Qnur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Publication:Communications of the ACM*, 37(1):81–88, 2016.

[96] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19, 2017.

[97] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. What makes users trust a chatbot for customer service? an exploratory interview study. In *International conference on internet science*, pages 194–208. Springer, 2018.

[98] Ankur Gandhe, Ariya Rastrow, and Bjorn Hoffmeister. Scalable language model adaptation for spoken dialogue systems. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 907–912. IEEE, 2018.

[99] Kai Gao, Hua Xu, and Jiushou Wang. A rule based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, 42:4517–4528, 2015.

[100] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

[101] Robert W. Gehl. Teaching to the turing test with cleverbot. *The Journal of Inclusive Scholarship and Pedagogy*, 24:56–66, 2014.

[102] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. A knowledge-grounded neural conversation model. In *AAAI*, 2018.

[103] Eun Go and S. Shyam Sundar. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, 97:304–316, 2019.

[104] Google. Word2vec googlenews binary database.

[105] googlecom. Building and deploying a chatbot by using dialogflow (overview). https://cloud.google.com/solutions/building-and-deploying-chatbot-dialogflow.

[106] Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Frank Enos, Julia Hirschberg, and Sachin Kajarekar. Combining prosodic lexical and cepstral systems for deceptive speech detection. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

[107] Arthur C Graesser, Katja Wiemer-Hastings, Peter Wiemer-Hastings, Roger Kreuz, Tutoring Research Group, et al. Autotutor: A simulation of a human tutor. *Cognitive Systems Research*, 1(1):35–51, 1999.

[108] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 2005.

[109] David Griol, Javier Carbó, and José M. Molina. An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence*, 27(9):759–780, 2013.

[110] Z. Guan, L.Chen, W. Zhao, Y. Zheng, S. Tan, and D. Cai. Weakly-supervised deep learning for customer review sentiment classification. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, 2016.

[111] C. Guggilla, T. Miller, and I.Gurevych. Cnn-and lstm-based claim classification in online user comments. In *In Proceedings of the International Conference on Computational Linguistics*, 2016.

[112] Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Event-Driven Emotion Cause Extraction with Corpus Construction*, 2016.

[113] Fenfei Guo, Angeliki Metallinou, Chandra Khatri, Anirudh Raju, Anu Venkatesh, and Ashwin Ram. Topic-based evaluation for conversational bots. *arXiv preprint arXiv:1801.03622*, 2018.

[114] Tager-Flusberg H. Evaluating the theory-of-mind hypothesis of autism. *Current Directions in Psychological Science*, 16:311–315, 2007.

[115] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online, 2020. Association for Computational Linguistics.

[116] K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of INTERSPEECH ISCA Singapore*, 2014.

[117] Jan-Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. Approaches for dialog management in conversational agents. *IEEE Internet Computing*, 23(2):13–22, 2018.

[118] Shizhu He, Cao Liu, Kang Liu, and Jun Zhao. Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, page 199–208, 2017.

[119] Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In *SIGDIAL*, pages 292–299, 2014.

[120] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[121] Ho Thao Hien, Cuong Pham-Nguyen, Le Nguyen Hoai Nam, and Thang Le Dinh. Intelligent assistants in higher-education environments: The fit-ebot, a chatbot for administrative and learning support. In *ICPS*, 2018.

[122] Rob High. The era of cognitive systems: An inside look at ibm watson and how it works. *IBM Corporation, Redbooks*, 1:16, 2012.

[123] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49:245–250, 2015. https://www.sciencedirect.com/science/article/pii/S0747563215001247.

[124] Sebastian Hobert. Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program. *Digital Learning Environment and Future IS Curriculum*, 2019.

[125] Daniel Homburg, Mirja Sophie Thieme, Johannes Völker, and Ruth Stock. Robotalk-prototyping a humanoid robot as speech-to-sign language translator. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[126] Matthew B. Hoy. Human-aided bots. *Medical Reference Services Quarterly*, 37(1):81–88, 2018.

[127] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*, pages 946–956. Springer, 2019.

[128] Dorit Hutzler, Esther David, Mirray Avigal, and Rina Azoulay. Learning methods for rating the difficulty of reading comprehension questions. In *2014 IEEE International Conference on Software Science, Technology and Engineering*, 2014.

[129] Ismail L. I., Verhoeven T., Dambre J., and Wyffels F. Leveraging robotics research for children with autism: A review. *International Journal of Social Robotics*, pages 1–22, 2018.

[130] Nobuo Inui, Takuya Koiso, Junpei Nakamura, and Yoshiyuki Kotani. Fully corpus-based natural language dialogue system. In *Natural Language Generation in Spoken and Written Dialogue, AAAI Spring Symposium*, 2003.

[131] Corbett J. Special language and political correctness. *British foumal of Special Education*, 21:17–19, 1994.

[132] M. Jain, S. Narayan, P. Balaji A. Bhowmick, R.K. Muthu, K.P. Bharath, and R. Karthik. Speech emotion recognition using support vector machine.

[133] Michel Galley Jianfeng Gao and Lihong Li. Neural approaches to conversational ai. *arXiv preprint. arXiv*, arXiv:1809.08267, 2019.

[134] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics.

[135] Petersilia J.R. Crime victims with developmental disabilities: A review essay. *Criminal Justice and Behavior*, 2001.

[136] Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn Walker. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 152–162, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[137] J.Wang, L-C. Yu, R.K.Lai, and X.Zhang. Dimensional sentiment analysis using a regional cnn-lstm model. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2016.

[138] Sahak Kaghyan, Shubham Sarpal, Andrei Zorilescu, and David Akopian. Review of interactive communication systems for business-to-business (b2b) services. *Electronic Imaging*, 2018(6):117–1, 2018.

[139] Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, and Chandan K. Reddy. Deep reinforcement learning for sequence to sequence models. *arXiv preprint arXiv:1805.09461*, 2018.

[140] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *arXiv preprint arXiv:1708.05148*, 2017.

[141] A Kim, Hyun-Je Song, Seong-Bae Park, et al. A two-step neural dialog state tracker for task-oriented dialog processing. *Computational intelligence and neuroscience*, 2018, 2018.

[142] Jinhyeon Kim, Donghoon Ham, Jeong-Gwan Lee, and Kee-Eung Kim. End-to-end document-grounded conversation with encoder-decoder pre-trained language model. In *DSTC9 Workshop, AAAI*, 2021.

[143] Na-Young Kim, Yoonjung Cha, and Hea-Suk Kim. Future english learning: Chatbots and artificial intelligence. *Multimedia-Assisted Language Learning*, 22(3):32–53, 2019.

[144] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization.

[145] Carlos Delgado Kloos, Cristina Catálan, Pedro J Muñoz-Merino, and Carlos Alario-Hoyos. Design of a conversational agent as an educational tool. In *2018 Learning With MOOCS (LWMOOCS)*, pages 27–30. IEEE, 2018.

[146] Lorenz Cuno Klopfenstein, Saverio Delpriori, and Alessio Ricci. Adapting a conversational text generator for online chatbot messaging. In *International Conference on Internet Science*, pages 87–99. Springer, 2018.

[147] Bence Kollanyi. Automation, algorithms, and politics— where do bots come from? an analysis of bot codes shared on github. *International Journal of Communication*, 10:20, 2016.

[148] Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Ross Beveridge, Jaime Ruiz, Bruce Draper, et al. Communicating and acting: Understanding gesture in simulation semantics. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*, 2017.

[149] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[150] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

[151] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.

[152] Keeheon Lee, Jeongwon Jo, Jinyoung Kim, and Younah Kang. Can chatbots help reduce the workload of administrative officers?-implementing and deploying faq chatbot service in a university. In *International Conference on Human-Computer Interaction*, pages 348–354. Springer, 2019.

[153] Kyumin Lee, Brian Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[154] Sungjin Lee, Qi Zhu, Ryuichi Takanobu, Zheng Zhang, Yaoqin Zhang, Xiang Li, Jinchao Li, Baolin Peng, Xiujun Li, Minlie Huang, and Jianfeng Gao. ConvLab: Multi-domain end-to-end dialog system platform. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 64–69, Florence, Italy, 2019. Association for Computational Linguistics.

[155] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues, 2017. https://arxiv.org/abs/1706.05125.

[156] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[157] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint*, arXiv:1510.03055, 2015.

[158] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint*, arXiv:1603.06155, 2016.

[159] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. arXiv:1606.01541.

[160] L. Li, Y. Zhao, D. Jiang, and Y. Zhang etc. Hybrid deep neural network-hidden markov model (dnn-hmm) based speech emotion recognition. In *In Proceedings Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013.

[161] Toby Jia-Jun Li, Amos Azaria, and Brad A Myers. Sugilite: creating multimodal smartphone automation by demonstration. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 6038–6049, 2017.

[162] Xuan LI, Huixin ZHONG, Bin ZHANG, and Jiaming ZHANG. A general chinese chatbot based on deep learning and its' application for children with asd. *International Journal of Machine Learning and Computing*, pages 1–10, 2020.

[163] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017.

[164] Yuting Liao and Jiangen He. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *IUI 20*, pages 430–442, 2020.

[165] C. Liebeskind and S. Liebeskind. Identifying abusive comments in hebrew facebook. In *ICSEE*, 2018.

[166] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[167] Phoebe Lin, Jessica Van Brummelen, Galit Lukin, Randi Williams, and Cynthia Breazeal. Zhorai: Designing a conversational agent for children to explore machine learning concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13381–13388, 2020.

[168] Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *LREC-2016*, 2016.

[169] Bing Liu and Ian Lane. Iterative policy learning in end-to-end trainable task-oriented neural dialog models. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 482–489. IEEE, 2017.

[170] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study

of unsupervised evaluation metrics for dialogue response generation. arXiv preprint https://arxiv.org/abs/1603.08023.

[171] Huiting Liu, Tao Lin, Hanfei Sun, Weijian Lin, Chih-Wei Chang, Teng Zhong, and Alexander Rudnicky. Rubystar: A non-task-oriented mixture model dialog system. *arXiv preprint arXiv:1711.02781*, 2017.

[172] X. Liu, Q. Wu, W. Zhao, and X. Luo. Technology-facilitated diagnosis and treatment of individuals with autism spectrum disorder: an engineering perspective. *Appl Sci*, 7(10):1051, 2017.

[173] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. Talk to me: Exploring user interactions with the amazon alexa. *Journal of Librarianship and Information Science*, 51(4):984–997, 2019.

[174] Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*, 2017.

[175] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proc. SIGDIAL 16*, pages 285–294, 2017. arXiv preprint https://arxiv.org/abs/1506.08909.

[176] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, 38(6):937–947, 2019.

[177] L.Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis:a survey. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discover*, 2018.

[178] A. Maheshwari. Report on text classification using cnn, rnn & han, 2019.

[179] ANNA MARIA. Got an alexa? you've got a polyglot tutor that can teach you a language. https://www.fluentu.com/blog/can-alexa-teach-languages/.

[180] Maria das Graças Bruno Marietto, Rafael Varago de Aguiar, Gislene de Oliveira Barbosa, Wagner Tanaka Botelho, Edson Pimentel, Robson dos Santos França, and Vera Lúcia da Silva. Artificial intelligence markup language: a brief tutorial. *arXiv preprint arXiv:1307.3091*, 2013.

[181] M. Marom, L. Uziel, and N.Denise. Children with disabilities in risk situations:a literature review. *Social Security (Hebrew edition)*, 99, 2016.

[182] Julia Masche and Nguyen-Thinh Le. A review of technologies for conversational systems. In *International conference on computer science, applied mathematics and applications*, pages 212–225. Springer, 2017.

[183] Julia Masche and Nguyen-Thinh Le. A review of technologies for conversational systems. In *International Conference on Computer Science, Applied Mathematics and Applications ICCSAMA 2017*, pages 212–225, 2018.

[184] Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic. The semaine corpus of emotionally coloured character interactions. In *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, ICME*, pages 1–4, 2010.

[185] Saul McLeod. Maslow's hierarchy of needs. *Simply psychology*, 1(1-18), 2007.

[186] Michael McTear. The role of spoken dialogue in user–environment interaction. *Human-Centric Interfaces for Ambient Intelligence*, pages 225–254, 2010.

[187] Allouche Merav, Azaria Amos, Azoulay Rina, Ben-Izchak Ester, Zwilling Moti, and Zachor Ditza A. Automatic detection of insulting sentences in conversation. In *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, pages 1–4. IEEE, 2018.

[188] Chkroun Merav and Azaria Amos. "did i say something wrong?": Towards a safe collaborative chatbot. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[189] Chkroun Merav and Azaria Amos. Safebot: A safe collaborative chatbot. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[190] Raphael Meyer von Wolff, Sebastian Hobert, and Matthias Schumann. How may i help you?–state of the art and open research questions for chatbots at the digital workplace. In *Proceedings of the 52nd Hawaii international conference on system sciences*, 2019.

[191] M.Huang, Q. Qian, and X. Zhu. Encoding syntactic knowledge in neural networks for sentiment classification. *ACM Transactions on Information Systems*, 35, 2005.

[192] Tomas Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239. IEEE, 2012.

[193] Xu Min, Duan Ling-Yu, Cai Jianfei, Chia Liang-Tien, Xu Changsheng, and Tian Qi. Hmm-based audio keyword generation. In *Pacific-Rim Conference on Multimedia*, pages 566–574. Springer, 2004.

[194] MindMeld. Introducing mindmeld . https://www.mindmeld.com/docs/intro/introducing_mindmeld.html.

[195] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. Survey of conversational agents in health. *Expert Systems with Applications*, 129:56–67, 2019.

[196] Robert C Moore, John Dowding, Harry Bratt, Jean Mark Gawron, Yonael Gorfu, and Adam Cheyer. Commandtalk: A spoken-language interface for battlefield simulations. In *Fifth Conference on Applied Natural Language Processing*, pages 1–7, 1997.

[197] Nikola Mrksic, Diarmuid O Seaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. Neural belief tracker: Data-driven dialogue state tracking. In *ACL*, pages 1777–1788, 2017. https://doi.org/10.18653/v1/P17-1163.

[198] Tom Nadarzynski, Oliver Miles, Aimee Cowie, and Damien Ridge. Acceptability of artificial intelligence (ai)-led chatbot services in healthcare: A mixed-methods study. *Digital health*, 5, 2019.

[199] Roberto Navigli. Natural language understanding: Instructions for (present and future) use. In *Proc. of AAAI*, IJCAI'18, page 5697–5702, 2018.

[200] Mark A Neerincx, Willeke van Vught, Olivier Blanson Henkemans, Elettra Oleari, Joost Broekens, Rifca Peters, Frank Kaptein, Yiannis Demiris, Bernd Kiefer, Diego Fumagalli, et al. Socio-cognitive engineering of a robotic partner for child's diabetes self-management. *Frontiers in Robotics and AI*, 6:118, 2019.

[201] Anh Nguyen and Wayne Wobcke. An agent-based approach to dialogue management in personal assistants. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 137–144, 2005.

[202] Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences*, pages 38–52. Springer, 2017.

[203] Ketakee Nimavat and Tushar Champaneria. Chatbots: An overview types, architecture, tools and future possibilities. *Int. J. Sci. Res. Dev*, 5(7):1019–1024, 2017.

[204] N.Kalchbrenner, E. Grefenstette, and P.Blunsom. A convolutional neural network for modelling sentences. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2014.

[205] Mansbach Noa, Neiterman Evgeny Hershkovitch, and Azaria Amos. An agent for competing with humans in a deceptive game based on vocal cues. *Proc. Interspeech 2021*, pages 4134–4138, 2021.

[206] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[207] Vahid Noroozi, Yang Zhang, Evelina Bakhturina, and Tomasz Kornuta. A fast and robust bert-based dialogue state tracker for schema-guided dialogue dataset. *CoRR*, abs/2008.12335, 2020.

[208] Mohammad Nuruzzaman and Omar Khadeer Hussain. A survey on chatbot implementation in customer service industry through deep neural networks. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, pages 54–61. IEEE, 2018.

[209] Rajai Nuseibeh. What is a chatbot?, 2018. https://medium.com/rajai_nuseibeh/what-is-a-chatbot-402427354f44.

[210] T. Nwe, S. Foo, and L. De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623, 2003.

[211] Geurts P., Ernst D., and Wehenkel L. Extremely randomized trees,. *Machine Learning*, 63:3–42, 2006.

[212] José Paladines and Jaime Ramirez. A systematic literature review of intelligent tutoring systems with dialogue in natural language. *IEEE Access*, 8:164246–164267, 2020.

[213] Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *Journal of medical Internet research*, 21(4):e12887, 2019.

[214] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*, 2005.

[215] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[216] Kishore A Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. Understanding affective experiences with bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, 2002.

[217] Do Eun Park, Yee-Jin Shin, EunAh Park, In Ae Choi, Woo Yeon Song, and Jinwoo Kim. Designing a voice-bot to promote better mental health: Ux design for digital therapeutics on adhd patients. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[218] Fivian Pascal and Reiser Dominique. Speech classification using wav2vec 2.0, 2021. https://www.zhaw.ch/storage/engineering/institute-zentren/cai/BA21_Speech_Classification_Reiser_Fivian.pdf.

[219] Leo Natan Paschoal, Aliane Loureiro Krassmann, Felipe Becker Nunes, Myke Morais de Oliveira, Magda Bercht, Ellen Francine Barbosa, and Simone do Rocio Senger de Souza. A systematic identification of pedagogical conversational agents. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2020.

[220] Leo Natan Paschoal, Lucas Fernandes Turci, Tayana Uchôa Conte, and Simone RS Souza. Towards a conversational agent to support the software testing education. In *Proceedings of the XXXIII Brazilian Symposium on Software Engineering*, pages 57–66, 2019.

[221] Andreea Peca, Adriana Tapus, Amir Aly, Cristina Pop, Lavinia Jisa, Sebastian Pintea, Alina Rusu, and Daniel David. Exploratory study: Children's with autism awareness of being imitated by nao robot. *arXiv preprint arXiv:2003.03528*, 2020.

[222] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[223] Denis Peskov, Nancy Clarke, Jason Krone, and Brigi Fodor. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *EMNLP-IJCNLP*, 2019.

[224] Xuan Lam Pham, Thao Pham, Quynh Mai Nguyen, Thanh Huong Nguyen, and Thi Thu Huong Cao. Chatbot as an intelligent personal assistant for mobile language

learning. In *Proceedings of the 2018 2nd International Conference on Education and E-Learning*, pages 16–21, 2018.

[225] Aditya Pradana, Goh Ong Sing, and YJ Kumar. Sambot-intelligent conversational bot for interactive marketing with consumer-centric approach. *International Journal of Computer Information Systems and Industrial Management Applications*, 6(2014):265–275, 2017.

[226] Dolça Tellols Maite Lopez-Sanchez Inmaculada Rodríguez Pablo Almajano Anna Puig. Enhancing sentient embodied conversational agents with machine learning. *Pattern Recognition Letters*, 129:317–323, 2020.

[227] Q. Qian, M. Huang, J. Lei, and X. Zhu. Linguistically regularized lstm for sentiment classification. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017.

[228] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. Alime chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 498–503, 2017.

[229] Livingstone Steven R and Russo Frank A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[230] Nicole Radziwill and Morgan Benton. Evaluating quality of chatbots and intelligent conversational agents. *Software Quality Professional*, 19(3):25, 2017.

[231] Winkler Rainer, Hobert Sebastian, Antti Salovaara, Sollner Matthias, and Leimeister Jan Marco. *Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent*, page 1–14. Association for Computing Machinery, New York, NY, USA, 2020. https://doi.org/10.1145/3313831.3376781.

[232] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics.

[233] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[234] Kiran Ramesh, Surya Ravishankaran, Abhishek Joshi, and K Chandrasekaran. A survey of design techniques for conversational agents. In *2017 ICICCT Information, Communication and Computing Technology*, page 336–350, 2017.

[235] Bhavika R Ranoliya, Nidhi Raghuwanshi, and Sanjay Singh. Chatbot for university related faqs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1525–1530. IEEE, 2017.

[236] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: a new benchmark and dataset. arXiv preprint https://arxiv.org/abs/1811.00207.

[237] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multidomain conversational agents: The schema-guided dialogue dataset. arXiv preprint https://arxiv.org/abs/1909.05855.

[238] K. Ravi and V. Ravi. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015.

[239] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, UK, 2000.

[240] Philip Resnik. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 1992.

[241] Mirray Avigal Rina Azoulay, Esther David and Dorit Hutzler. *Adaptive Task Selection in Automated Educational Software: A Comparative Study*, chapter 7. Elsevier, 2021.

[242] Alan Ritter, Colin Cherry, and Bill Dolan. Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California, jun 2010. Association for Computational Linguistics. https://www.aclweb.org/anthology/N10-1020.

[243] Sonia Roccas, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo. The big five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28(6):789–801, 2002.

[244] Ardila Rosana, Branson Megan, Davis Kelly, Henretty Michael, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[245] Casey Ross and Ike Swetlitz. Ibm's watson supercomputer recommended 'unsafe and incorrect'cancer treatments, internal documents show. *Stat*, 25, 2018.

[246] R.Socher, B.Huval, C.D.Manning, and A.Y.Ng. Semantic compositionality through recursive matrix-vector spaces. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2012.

[247] R.Socher, J.Pennington, E.H.Huang, A.Y.Ng, and C.D.Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

[248] H. Hajj S. Shaheen, W. El-Hajj and S. Elbassuoni. Emotion recognition from text based on automatically generated rules. In *IEEE International Conference on Data Mining Workshop*, 2014.

[249] Amir Sadeghipour and Stefan Kopp. Embodied gesture processing: Motor-based integration of perception and action in social artificial agents. *Cognitive computation*, 3(3):419–435, 2011.

[250] C.N.dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis for short texts. In *In Proceedings of the International Conference on Computational Linguistics*, 2014.

[251] Eric Saund. How do conversational agents answer questions? https://towardsdatascience.com/how-do-conversational-agents-answer-questions-d504d37ef1cc.

[252] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. Let's talk about race: Identity, chatbots, and ai. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[253] Van Gemert-Pijnen JE Scholten MR, Kelders SM. Self-guided web-based interventions: Scoping review on user needs and the potential of embodied conversational agents to address them. *Journal of medical Internet research*, 19(211), 2017.

[254] Nicolas Schrading, Cecilia Ovesdotter Alm, Ray Ptucha, and Christopher Homan. An analysis of domestic abuse discourse on Reddit. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2577–2583, Lisbon, Portugal, sep 2015. Association for Computational Linguistics. https://www.aclweb.org/anthology/D15-1309.

[255] Ryan M Schuetzler, G Mark Grimes, Justin Scott Giboney, and Jay F Nunamaker Jr. The influence of conversational agents on socially desirable responding. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, page 283, 2018.

[256] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[257] Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*, 2017.

[258] Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, Sai Rajeshwar, Alexandre de Brebisson, Jose M. R. Sotelo, Dendi Suhubdy, Vincent Michalski, Alexandre Nguyen, Joelle Pineau, and Yoshua Bengio. A deep reinforcement learning chatbot. arXiv:1709.02349.

[259] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems. https://arxiv.org/pdf/1512.05742.pdf.

[260] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *arXiv preprint arXiv:2101.07714*, 2021.

[261] Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. Reinforcement learning for spoken dialogue systems. In *Nips*, pages 956–962, 1999.

[262] Shekhar Singh, Akshat Jain, and Deepak Kumar. Recognizing and interpreting sign language gesture for human robot interaction. *International Journal of Computer Applications*, 52(11), 2012.

[263] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. https://arxiv.org/abs/2004.08449.

[264] R. Socher, A.Perelygin, J.Y.Wu, J.Chuang, C.D.Manning, A.Y.Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment tree bank. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013.

[265] Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, 2015. Association for Computational Linguistics. https://www.aclweb.org/anthology/N15-1020.

[266] Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64. Association for Computational Linguistics, 2020.

[267] Amanda Stent, John Dowding, Jean Mark Gawron, Elizabeth Owen Bratt, and Robert C Moore. The commandtalk spoken dialogue system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 1999.

[268] Donald J Stoner, Louis Ford, and Mark Ricci. Simulating military radio communications using speech recognition and chat-bot technology. *The Titan Corporation, Orlando*, 2004.

[269] Eliza Strickland. Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care. *IEEE Spectrum*, 56(4):24–31, 2019.

[270] Pei-Hao Su, David Vandyke, Milica Gasic, Dongho Kim, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. *arXiv preprint arXiv:1508.03386*, 2015.

[271] Venkatramanan S Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.

[272] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[273] Alexa Prize Taskbot. Alexa prize taskbot, 2021. https://developer.amazon.com/alexaprize.

[274] Z. Teng, D-T. Vo, and Y. Zhang. Context-sensitive lexicon features for neural sentiment analysis. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

[275] M.F. Tennyson, D.A. Kuester, J. Casteel, and C. Nikolopoulos. Accessible robots for improving social skills of individuals with autism. *J. Artif. Intell. Soft Comput.*, 21:267–277, 2016.

[276] Florian v Wangenheim Theresa Schachner, Roman Keller. Artificial intelligence-based conversational agents for chronic conditions: Systematic literature review. *Journal of Medical Internet Research*, 22, 2020.

[277] NT Thomas. An e-business chatbot using aiml and lsa. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2740–2742. IEEE, 2016.

[278] Jörg Tiedemann. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248, 2009.

[279] Van-Khanh Tran, Le-Minh Nguyen, and Satoshi Tojo. Neural-based natural language generation in dialogue using RNN encoder-decoder with semantic aggregation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 231–240, Saarbrücken, Germany, 2017. Association for Computational Linguistics.

[280] Charikleia Triantafyllidou. Assistive technologies for dyslexia: Punctuation and its interfaces with speech. Master's thesis, University of Central Florida, 2020.

[281] Maomi Ueno and Yoshimitsu Miyazawa. Irt-based adaptive hints to scaffold learning in programming. *IEEE Transactions on Learning Technologies*, 11(4):415–428, 2017.

[282] Bono V., Narzisi A., Jouen AL, Tilmont E., Hommel S.and Jamal W., Xavier J., Billeci L.and Maharatna K., Wald M., Chetouani M., Cohen D., Muratori F., and MICHELANGELO Study Group. Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (asd). *frontiers in psychiatry*, 2016.

[283] C. T. Valadão, C. Goulart, H. Rivera, E. Caldeira, T. F. Bastos Filho, A. Frizera-Neto, and R. Carelli. Analysis of the use of a robot to improve social skills in children with autism spectrum disorder. *Res Biomed Eng*, 32:161–175, 2016.

[284] Kees van Deemter, Emiel Krahmer, and Mariët Theune. Squibs and discussions: Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31(1):15–24, 2005.

[285] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[286] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. On evaluating and comparing conversational agents. *arXiv preprint arXiv:1801.03625*, 4:60–68, 2018.

[287] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju1. On evaluating and comparing conversational agents. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[288] Laduram Vishnoi. Conversational agent: A more assertive form of chatbots, 2020. https://towardsdatascience.com/conversational-agent-a-more-assertive-form-of-chatbots-de6f1c8da8dd.

[289] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. Developing a personality model for speech-based conversational agents using the psycholexical approach. In *CHI 2020*, pages 1–14, 2020.

[290] Richard S Wallace. The anatomy of alice. In *Parsing the turing test*, pages 181–210. Springer, 2009.

[291] X. Wang, W. Jiang, and Z. Luo. Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *In Proceedings of the International Conference on Computational Linguistics*, 2016.

[292] X. Wang, C. Sun Y. Liu, B. Wang, and X. Wang. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2015.

[293] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9.1:36–45, 1966.

[294] Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[295] Jason D Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoffrey Zweig. Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (luis). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 159–161, 2015.

[296] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[297] Magdalena Wolska, Quoc Bao Vo, Dimitra Tsovaltzi, Ivana Kruijff-Korbayová, Elena Karagjosova, Helmut Horacek, Armin Fiedler, and Christoph Benzmüller. An annotated corpus of tutorial dialogs on mathematical theorem proving. In *LREC*, 2004.

[298] Wood, L.J., Zaraki, A., Robins, B., Dautenhahn, and K. Developing kaspar: A humanoid robot for children with autism. *International Journal of Social Robotics*, 2019.

[299] Wu, T.H. Falk, and W. Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53:768–785, 2011.

[300] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510, 2017.

[301] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, , and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *AAAI*, volume 33, page 7346–7353, 2019.

[302] Özge Nilay Yalçın. Empathy framework for embodied conversational agents. *Cognitive Systems Research*, 59:123–132, 2020.

[303] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[304] Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, and Zhou Yu et al. On the generation of medical dialogues for covid-19. arXiv preprint https://arxiv.org/abs/2005.05442.

[305] Xi Yang, Marco Aurisicchio, and Weston Baxter. Understanding affective experiences with conversational agents. In *CHI 2019*, 2019.

[306] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint*, arXiv:1906.08237, 2019.

[307] Zi Yin, Keng-hao Chang, and Ruofei Zhang. Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2131–2139, 2017.

[308] Zhu Youxiang, Obyat Abdelrahman, Liang Xiaohui, Batsis John A, and Roth Robert M. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. *Proc. Interspeech 2021*, pages 3790–3794, 2021.

[309] Dong Yu and Li Deng. *AUTOMATIC SPEECH RECOGNITION.* Springer, 2016.

[310] J. Yu and J. Jiang. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

[311] Yanchao Yu, Arash Eshghi, Gregory Mills, and Oliver Lemon. The burchak corpus: a challenge data set for interactive learning of visually grounded word meanings. In *Proceedings of the Sixth Workshop on Vision and Language)*, pages 1–10. Association for Computational Linguistics, 2017.

[312] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency. Tensor fusion network for multimodal sentiment analysis. *Empirical Methods Natural Language Processing*, pages 1114–1125, 2017.

[313] Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21), 2020.

[314] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online, 2020. Association for Computational Linguistics.

[315] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too?, 2018. arXiv preprint https://arxiv.org/abs/1709.02349.

[316] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt : Large-scale generative pre-training for conversational response generation. arXiv preprint https://arxiv.org/abs/1911.00536.

[317] Zheng Zhang, Ryuichi Takanobu, Minlie Huang, and Xiaoyan Zhu. Recent advances and challenges in task-oriented dialog system. *CoRR*, abs/2003.07490, 2020.

[318] Tiancheng Zhao and Maxine Eskenazi. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. *arXiv preprint arXiv:1606.02560*, 2016.

[319] Z. Zhao, H. Lu, D. Cai, X. He, and Y.Zhuang. Microblog sentiment classification via recurrent random walk network learning. In *In Proceedings of the Internal Joint Conference on Artificial Intelligence*, 2017.

[320] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[321] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

[322] Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. *arXiv preprint arXiv:2002.04793*, 2020.

תקציר

ילדים ומבוגרים עם צרכים מיוחדים עשויים להתקשות בתקשורת עם הסביבה ובין היתר בזיהוי מצבים חברתיים מורכבים וסכנות. לפיכך, הם עשויים למצוא את עצמם מעליבים בלא כוונה את הסובבים אותם או להיות קורבן לניצול ואלימות.

מטרת המחקר הזה היא לעזור לילדים אלו להבין את הסביבה שלהם ולעזור ביצירת אינטראקציה נכונה איתה.

בעבודה זו אנו מציעים לפתח סוכן אוטונומי שיגלה מצבים חברתיים בעיתיים אלו ויאותת למשתמש :על ידי טקסט, דיבור או צורות איתות אחרות. כמו כן הוא יציע לילד כיצד להגיב.

המחקר שלנו מהווה אבן פינה לסוכן אוטומטי שיאפשר עזרה וייעוץ מקוון לילדים עם קשיי תקשורת ,באופן שיאפשר להם לתפקד טוב יותר בחברה באמצעות סיוע זה.

בעבודה זו היו שני שלבים:

בשלב הראשון, התחלנו עם ניתוח טקסט. בנינו מאגר נתונים עם 13490 משפטים כתובים שמסווגים לארבע קטגוריות: משפט "רגיל", משפט פוגע, משפט שלילי על אנשים בגוף שלישי, משפט שמעיד על מצב מסוכן שמצריך התערבות מידית.

חילקנו את מאגר הנתונים ביחס של 90% - 10%. השתמשנו בשיטות של למידת מכונה על מנת ללמוד מ 90% משפטים שנבחרו רנדומלית על 10% המשפטים האחרים.

הגענו לאחוזי הצלחה שקרובים ל 70%.

בשלב השני, בנינו מאגר נתונים שמורכב ממעל 2600 משפטים: קול וטקסט.

המשפטים מסווגים לשלוש קטגוריות: משפטים נטרליים, משפטים פוגעניים, ומשפטים שמעידים על מצבים מסוכנים ומצריכים התערבות מיידית.

גם כאן השתמשנו בשיטות למידת מכונה שונות, ושילבנו בין וקטורים של BERT ורשתות נוירוניות.

בעצם העבודה שלנו מעידה שניתן לבנות כזה סוכן שיעזור לילדים עם צרכים מיוחדים ויזהה מצבים פוגעניים ומצבים מסוכנים להם.

כמו כן כתבנו מאמר סקירה בנושא : סוכני שיחה (conversational agents) שבשנים האחרונות הפכו להיות יותר ויותר נוכחים בחיים שלנו. הצגנו את השימושים השונים שלהם, והטכנולוגיות השונות בהן הם פותחו.

בס"ד

אוניברסיטת אריאל בשומרון

סיוע לילדים עם צרכים מיוחדים בהבנת אינטראקציות חברתיות

עבודת מחקר זו הוגשה כחלק מדרישות התואר

"דוקטור לפילוסופיה"

מאת

מירב אלוש

עבודה זו נכתבה בהנחיית

פרופ' ודים לויט

פרפ' עמוס עזריה

דר' רינה אזולאי

יוני 2022