eDEM-CONNECT: An Ontology-Based Chatbot for Family Caregivers of People with Dementia

Maurice Boiting[1(✉)], Niklas Tschorn[1], Sumaiya Suravee[2], Kristina Yordanova[2], Margareta Halek[3], Franziska A. Jagoda[3], Stefan Lüdtke[4], and Anja Burmann[1]

[1] Fraunhofer Institute for Software and Systems Engineering ISST, 44147 Dortmund, Germany
{maurice.boiting,niklas.tschorn,anja.burmann}@isst.fraunhofer.de
[2] Institute for Data Science, Universität Greifswald, 17489 Greifswald, Germany
{sumaiya.suravee,kristina.yordanova}@uni-greifswald.de
[3] School of Nursing Science, Faculty of Health, Universiät Witten/Herdecke, 58453 Witten, Germany
{margareta.halek,franziska.jagoda}@uni-wh.de
[4] Institute for Visual and Analytic Computing, Universät Rostock, 18059 Rostock, Germany
stefan.luedtke@uni-rostock.de

Abstract. Home care of people with dementia (PwD) is mainly organized and carried out by non-professional family caregivers, who struggle to interpret the needs of PwD correctly and are confronted with the challenging behavior of their relatives. Although support services for family caregivers are widespread in Germany, they are rarely used due to the fact that information is poorly organized and relatives are faced with a flood of disorganized, outdated, and confusing content. Due to the technical development of chatbot technologies (*ChatGPT*), chatbots gain more and more relevance. Based on the new technological possibilities, we developed an online communication and service platform with an integrated chatbot within the *eDEM-CONNECT* project, with the aim of making structured and easily understandable information accessible for family caregivers. This work focuses on the development of a chatbot pipeline that has broad domain knowledge through a provided ontology on the topic of agitation of PwD. This allows the chatbot to provide relevant and peer-reviewed information to family members. In our approach, a patient history is first taken based on several diagnostic questions so that relevant information can be output in a later step. For this purpose, we demonstrate that agitations in natural language can be correctly recognized by the used *BERT model* and that our developed chatbot is able to select further diagnostic questions based on the predictions of a *Markov logic network*.

Keywords: chatbot · ontology · home care · natural language processing · transformer models · Markov logic network

M. Boiting and N. Tschorn—contributed equally to this work.

A. Cunha et al. (Eds.): MobiHealth 2023, LNICST 578, pp. 73–91, 2024.
https://doi.org/10.1007/978-3-031-60665-6_6

1 Introduction

The home care of people with dementia (PwD) is mostly organized and executed by their relatives. According to the German Ageing Survey (DEAS) 04/2022 [13], there are about 1.8 million people with dementia and a total of 1.4 million caregiving relatives of PwD in Germany in 2022 [9]. The main challenge for family caregivers in this context is the confrontation with agitation and challenging behavior [27]. Relatives are often no professional caregivers and struggle to recognize the needs of PwD correctly. PwD often express their needs in actions [18] that are misinterpreted as aggressive behavior [23]. The challenging behavior often results in an increasingly unstable relationship between the PwD and their caregiving relatives and therefore mostly leads to the PwD moving to a retirement home [8].

Help services for family caregivers are widespread in Germany but rarely used [16] due to the sheer amount of these services and regarding the fact that the information on them is mostly very poorly organized. Most users first contact occurs via an online search engine like Google. In such a case, the information is commonly presented unsorted and might even contain sources that contradict each other. Unhelpful and confusing content like outdated information or local law differences due to Germany's federal state system can not be filtered out. Caregivers have often expressed their need to understand the changing behavior of the PwD, emphasizing that, in their belief, the family home was the best place for the PwD to stay [19]. Therefore, there is a strong demand for user-centered, tailored and up to date information in the domain of familycare of PwD.

Studies have shown the potential that chatbots can provide in the area of home care of PwD. They offer constant availability and intuitive usability, especially for elderly people who tend to have a lower affinity towards current technology [17,28]. Nonetheless, chatbots that address such topics are still in their infancy and users are challenged when trying to solve more complex problems with these bots [22].

In recent years, chatbot technology has gained an increased amount of attention in the technological and scientific field. More and more schematic tasks, typically performed by humans (like telephone services), are now outsourced to chatbots [24]. Due to its rising popularity in recent years, chatbot technology has improved as well, enabling the bots to handle increasingly complex tasks [24]. New use cases keep occurring and raise the need for new, more advanced and problem-specific chatbot technology. Easy tasks might just require a simple question-answer pattern, while more complex tasks require a set of multiple questions, the need to store the conversation context during the runtime of the bot or the domain knowledge of complex thematical landscapes. Designing such technology for more advanced user scenarios to be addressed by chatbots is a relevant question in the current chatbot development and scientific research [5].

In the *eDEM-CONNECT* project we developed a chatbot-based online communication and service platform. For intelligent interactions of the chatbot with caregiving relatives, we present a new approach based on a transformer model

for understanding user input and a *Markov logic network (MLN)* to be able to react intelligently to these user concerns through reasoning.

The goal of this project was to provide structured information for family caregivers on dementia and agitation, empowering them to confront their everyday challenges with their relatives. In contrast to the unstructured nature of search engines, the information should be presented in an easy to understand and user-friendly form. Besides information on dementia and agitation itself, the platform provides information on local care facilities, which can be categorized and searched based on their offered services. The user can utilize the chatbot to navigate through the website and also communicate directly with the bot to find the exact help texts and instructions he needs in a given situation. To ensure the quality and correctness of the provided information, the chatbot uses an expert-validated ontology as its knowledge base, in contrast to other chatbots that only operate on raw dialog data. This is crucial due to the fact that false information in a health-critical domain like this one can possibly cause great harm.

In this paper, we present the results of the project *eDEM-CONNECT* and discuss whether a chatbot is able to offer a helpful user interface and functionality in this specific context. The chatbot is operated in German language and provides curated and purposeful information instead of the more chaotic results of search engines. It offers an intuitive and easy to use interface, even for elderly users with low technical affinity. We examine how effectively the chatbot is capable to identify the domain-specific concepts and user intents and discuss the potential for further development and technical limitations.

2 Related Work

A chatbot is a computer program designed to hold an intelligent conversation between a human user and the bot itself [1]. The most simple chatbots come in the form of decision trees, while more complex chatbots are able to have open conversations with the user. They rely on full-text search engines, searching for specific keywords to identify the users' intents [1,2].

A current milestone in the development of chatbot technology is the project *ChatGPT* by OpenAI. It features a non-topic-specific chatbot that is able to react and chat about a wide variety of subjects and is even able to perform creative tasks like generating a speech on a given topic. *ChatGPT* was launched in November 2022 and had already reached a number of 100 million active users in January 2023, only two months after its release [7,14].

In the medicine domain, the diagnostic app *Ada Health* enables the user to specify his symptoms, which are then analyzed by its chatbot. The bot provides possible causes of the symptoms and also a likeliness parameter for the given diagnoses or the option to consult a doctor directly.[1] *OneRemission* is an app designated to cancer patients and survivors. Its chatbot provides diets, exercises and post-cancer activities, empowering the user's independence. The

[1] https://ada.com/de/ (accessed: 2023-11-01).

provided information is curated by medical experts, so the user does not always need to rely on a doctor regarding a cancer-specific question.[2] *MediBot* uses an integrated ontology as its knowledge base and provides information on drug medication for Portuguese-speaking users [4]. *DigiCare* combines natural language processing techniques and dynamic Bayesian inference to provide a conversational intelligent tutoring system, aiming to deliver study materials for nursing subjects [25].

In the dementia and care domain, *Elisabot* is a chatbot that supports its users in reminiscence therapy by showing them pictures of past experiences and asking therapy questions about them. The questions regarding the pictures are automatically generated by the chatbot [6]. *AlzBot* offers training for the user to challenge the memory loss of Alzheimer's patients and also provides the opportunity to track the location of the patient to reduce the caregivers' burden [15]. The *Androz Chatbot* and *Companion Chatbot* also provide memory training but also offer features like geo-fencing that offer guidance for PwD that struggle with orientation when wandering around [26,29].

The *Care* chatbot developed by Müller et al. [17] provided a set of questionnaires, mostly related to biographic contents. In a study they examined the potential use cases and effectiveness of chatbots in the dementia domain. In conclusion, the chatbots were welcomed by the users and provided information, empowering the patients to be more self-reliant, but on the other hand, marked the current limitations and also pointed out the importance of human caregivers that will still remain irreplaceable. Due to the fact that most chatbots are still in their infancy, the use cases remain mostly simple tasks, and more complex user scenarios offer potential for further development.

3 Methods

In chatbot development, several questions have to be addressed. First of all, the exact problem of the caregiver concerning the PwD must be identified to provide suitable knowledge. In addition, depending on the current course of the conversation, a chatbot needs to ask other appropriate questions in order to ensure an optimal diagnosis. In our approach, we addressed this problem through different methods and present a chatbot pipeline that serves as a support for relatives of PwD. Our pipeline is therefore divided into three parts (cf. Fig. 1): interpretation of the user input with the *BERT model (Bidirectional Encoder Representations from Transformers)* [10], prediction of the next diagnosis question with a *Markov logic network (MLN)* and output of relevant knowledge with a given ontology created by domain experts. In the next sections, we will first present the conversation flow with the chatbot and then introduce our ontology-based chatbot pipeline.

[2] https://keenethics.com/project-one-remission (accessed: 2023-11-01).

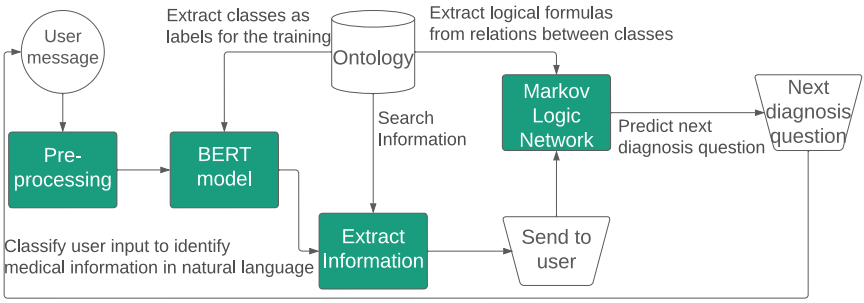


Fig. 1. The figure shows the technical pipeline of the chatbot. First, the user messages are preprocessed, removing umlauts, punctuation (except question marks), double spaces, pronouns and names of reference persons. In addition, only lowercase letters were used. Preprocessed user messages are classified according to medical ontology classes using the BERT model. Based on these identified classes, queries are used to extract information from the ontology that is forwarded to the user as the response. Finally, to ask further diagnostic questions, a Markov logic network is used to predict the most relevant diagnostic question for the current situation.

3.1 Chatbot Conversation Flow

In the course of a chat conversation, the chatbot asks several diagnostic questions until the problem has been sufficiently localized and advice can be given to solve the relative's problem. For this purpose, the chatbot is able to ask open and closed questions, where the closed questions can only be answered with *yes*, *no* or *maybe*.

This mixture of open as well as closed questions allows the chatbot to effectively recognize medical concepts in the user's chat messages. For each recognized concept, help texts with concrete suggestions for solving the problem (e.g., tips for communication or guidelines to recognize pain in PwD) are provided.

The user can choose for each help text whether to access a long or a short version of the provided content. Furthermore, in certain cases (e.g., more urgent problems), intervention strategies are issued in addition to the help texts to give the relatives concrete recommendations for action. A schematic representation of the conversation flow is shown in Fig. 2.

In addition to interpreting answers to diagnostic questions, the chatbot is also able to recognize other user intents (cf. Table 1) such as asking for regional help or requesting completely different topics.

3.2 Ontology Driven Knowledge Base

In order to provide caregivers helpful knowledge and information such as help texts, advice, intervention strategies as well as contact details of local institutions, it is necessary that the chatbot has a broad domain knowledge in a

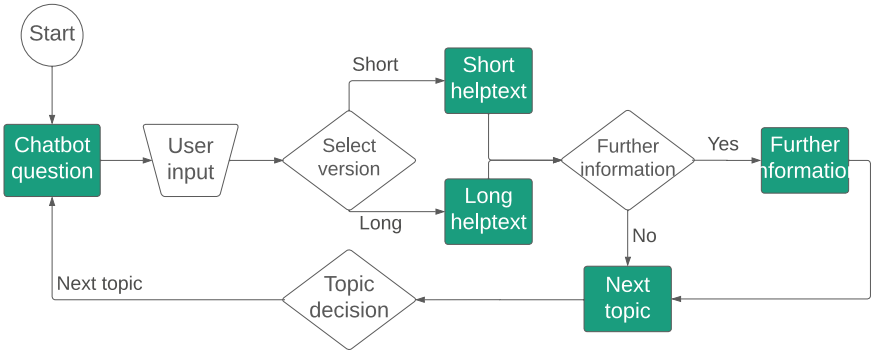


Fig. 2. The figure shows the flow of the conversation with the chatbot. The chatbot first asks a diagnosis question, which the user answers. A help text is then given to the user in order to be able to provide assistance with the user's problem. Since help texts are available in a long and a short version, the user is asked beforehand whether the short or the full long version should be displayed. Afterwards, regardless of the version selected, the user is asked whether further information (e.g., in the form of links) should be delivered. Finally, the user is asked on which topic he would like to have information next in order to be able to ask another diagnosis question.

machine-readable representation. Therefore, our approach uses a topic-specific ontology on the challenging behavior of PwD. An ontology is described as a structured representation of knowledge, encompassing a set of concepts situated in a specific domain, along with the relationship that is linked between these concepts [12].[3]

In our case, different aspects associated with dementia were modeled in the ontology by domain experts (cf. Fig. 3): The *agitation* i.e., the behavior of the PwD, *causes* that can cause the agitation, *consequences* caused by the agitation, characteristics of the *person with dementia* itself (e.g., abilities and medication) and intervention strategies (*interventions*).

In order to narrow down the problem, the approach was limited to only five different types of agitation:

- *Physical aggressive (PA)*: e.g., the PwD is throwing objects or has violent behavior,
- *Physical non-aggressive (PNA)*: e.g., the PwD is wandering around at night or is restless,
- *Verbal aggressive (VA)*: e.g., the PwD offends other people,
- *Verbal non-aggressive (VNA)*: e.g., the PwD keeps repeating the same question,
- *Resisting care (RC)*: e.g., the PwD refuses to be washed or dressed.

[3] The *eDEM-CONNECT* ontology was developed in the Web Ontology Language (OWL) using the software *Protégé*: https://protege.stanford.edu/ (accessed: 2023-08-31).

Table 1. Overview of all chatbot intents and representation of corresponding utterances.

Intent	Description	Example user utterances
Answer to diagnosis question	The user answers a question posed by the chatbot	Yes, my husband is in pain
Request for another topic	The user wants information on a completely different topic	I would rather know more about communication
Local help	Provides an overview of aids in the user's local area	Where can I get help near me?
Change helptext version	Changes the available version of the last sent help text	I would prefer to have the long version

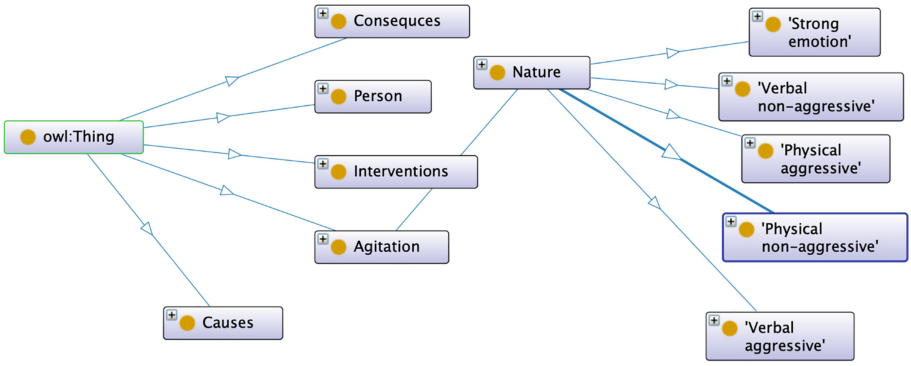


Fig. 3. An excerpt of the classes from the ontology that are relevant for the chatbot. Hierarchical relationships between the concepts are shown. The figure was taken as a screenshot from the software *Protégé*.

Helpful and unknown to the user information is searched in the ontology via *SPARQL*[4] queries, based on the previous conversation history as well as the latest user message. The SPARQL query therefore retrieves all information collected about the situation (e.g., the agitation description of the relative) as input and enables searching the ontology for relevant help texts or other information. The query result is then forwarded to the user as a chat message.

We formally define $\mathbf{\Gamma}$ as the current conversational context that determines which answer or diagnostic question is passed next to the user. Here, the context $\mathbf{\Gamma}$ characterizes the identified problem as a set of nodes ω from the ontology $\mathbf{\Omega}$ that have been marked as applicable to the situation during the course of the conversation.

Since diagnostic questions can be answered with *yes*, *no* or *maybe*, it is necessary for the chatbot to store the identified ontology nodes with different degrees

[4] https://www.w3.org/TR/sparql11-query/ (accessed: 2023-08-31).

of knowledge during the course of the conversation. Due to the different knowledge levels, it is therefore possible to output different help texts for different situations. For example, if the family member is not sure whether the person being cared for is in pain, a help text on the topic of pain recognition can be offered first.

3.3 Interpretation of the User Input

An important step for the communication between chatbot and caregiver is the understanding of the specific problem described in natural language. For the purpose of constraining the problem as well as for further interpretation, we developed a mapping from a chat message $x \in \mathbb{X}$ to a set of medical concepts $c \in \mathbf{\Omega}$ and the other defined chatbot intents (e.g., request for local help):

$$f : \mathbb{X} \rightarrow \mathbf{\Omega} \cup \{\text{Chatbot Intents}\} \tag{1}$$

In our approach, we use the BERT model [10] to classify text messages along classes $c \in \mathbf{\Omega}$ from the ontology $\mathbf{\Omega}$ and other intents. BERT relies on multiple Transformer-Blocks to capture relationships between words and sentences with a so-called attention mechanism [30]. Due to the fact that several concepts are often mentioned in one sentence (e.g., the relative describes that the family member is loud and restless), the presented model follows a multi-label approach. This allows multiple concepts from the ontology to be recognized in one chat message.

Data. For the demonstration of this approach, we use seven concepts from our ontology as labels for the training: *Physical aggressive (PA)*, *Physical non-aggressive (PNA)*, *Verbal aggressive (VA)*, *Verbal non-aggressive (VNA)* and *Resisting care (RC)*. To answer yes, no and maybe questions, additional concepts for *Yes*, *No* and *Unsure* were added. In order to enable further interactions with the chatbot as described in Sect. 3.1, we added multiple intents to the concept set, such as *Short version*, *Long version* and *Local help*. Furthermore, a rejection class *None* was added to allow the model to select none of the ontology concepts. For this purpose, several sentences about non-medical topics were randomly collected from the English *Wikipedia* and added to the training data set.

The training data set regarding the other concepts was created by experts from the nursing domain by considering possible example sentences users could ask the chatbot. A total of 407 samples were collected and divided into five random folds for the k-fold cross-validation procedure. A similar class distribution was obtained for each fold by stratification. An overview of the total samples per class is shown in Fig. 4.

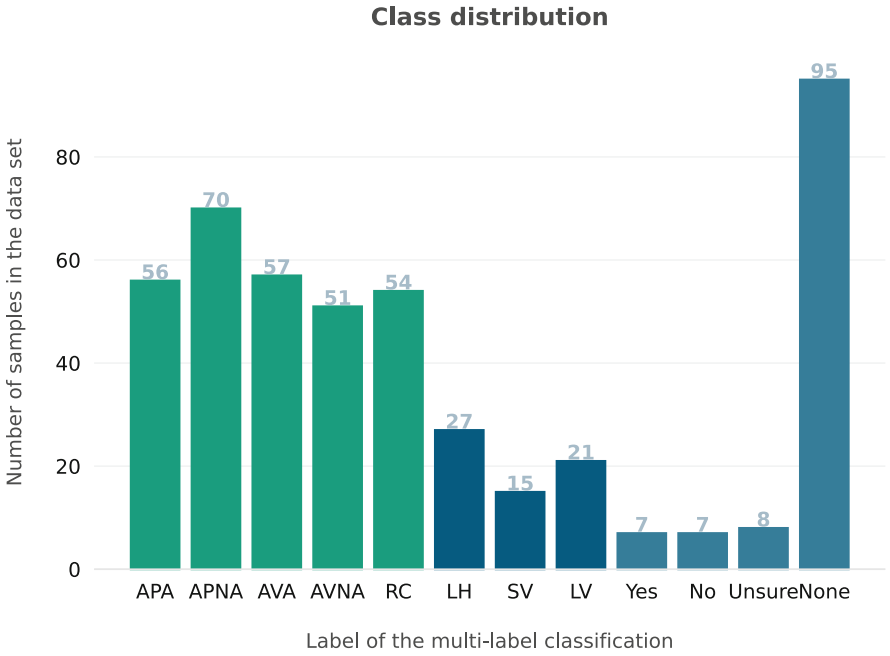


Fig. 4. The graph shows the number of samples used per class. Agitations in green: *Physical aggressive (PA)*, *Physical non-aggressive (PNA)*, *Verbal aggressive (VA)*, *Verbal non-aggressive (VNA)* and *Resisting care (RC)*; intents in blue: *Local help (LH)*, *Short version (SV)* and *Long version (LV)* and *Yes*, *No* and *Unsure* in light blue. (Color figure online)

Preprocessing and Data Augmentation. Each sample is preprocessed in several steps: Umlauts, punctuations (except for question marks, as these could imply a question) and double spaces were removed. In addition, only lowercase letters were used. To avoid the model learning a gender-specific prediction, all terms such as pronouns and names of reference persons such as husband, mother, etc. were removed and replaced by the token *PwD* (person with dementia).

Due to the small amount of training data, the training data was further augmented using back-translation, which results in a total data set of 1.928 samples. In this method, the sample chatbot queries in German language are first translated into several other languages and then back into the original language [11].

Training. In order to learn the mapping $f : \mathbb{X} \rightarrow \mathbf{\Omega} \cup \{\text{Chatbot Intents}\}$ (see Eq. 1) from natural language to medical concepts and other intents, the large language model BERT pre-trained on German was fine-tuned within 100 epochs using our generated data set.

The BERT model [10] uses a bidirectional approach, processing input from both left to right and right to left. Therefore, the model is able to understand

the context of a phrase considering the previous and subsequent words, resulting in a deeper semantic representation. As a result, the model is able to generate contextual word representations, which allows the model to develop a better understanding of the relationship between words and phrases, making it ideally suited for comprehension tasks such as text classification[5].

3.4 Prediction of the Next Diagnosis Question

To offer suitable questions and content for further diagnosis depending on the current conversation, a mapping from the conversation context to classes from the ontology is necessary. These classes are linked to diagnostic questions as well as other content. Thus, a prediction of the next ontology class describes the prediction of the next relevant diagnosis question that is most likely to apply to the situation and be most helpful in solving the relative's problem.

Therefore, we have developed a mapping to predict the next most likely class $\omega \in \mathbf{\Omega}$ from the ontology $\mathbf{\Omega}$ based on the previous conversational $\mathbf{\Gamma}$, which represents the current conversational context:

$$g : \mathbf{\Gamma} \rightarrow \mathbf{\Omega} \tag{2}$$

For the prediction of the next relevant ontology class, we used a *Markov logic network (MLN)*. Markov logic networks combine first-order logic and probabilistic graphical models, enabling the modeling of complex domains with formulas consisting of variables and predicates [21]. Thus, contrary to the direct use of inference in the ontology, the possibility of mapping uncertainties is provided. For this purpose, we have modeled the information about the current situation as well as the domain knowledge as a set of predicates and logical formulas. The objective of the MLN is to determine a probability distribution about which ontology classes might be relevant for the user's problem.

We have modeled an MLN with three observable (knowledge we have gathered so far about the case) and one hidden predicate:

- *Know(X)*: The observable predicate *Know(X)* describes the knowledge gathered in the course of the conversation so far and therefore represents all ontology classes that are reliably applicable to the conversation (verified by asking various diagnostic questions). For example, the predicate *Know(Pain)* indicates that the chatbot knows that the PwD is in pain.
- *Maybe(X)*: The observable predicate *Maybe(X)* describes all the information we have not collected with full certainty about the situation. This is, for example, information that the relative has only indicated with "maybe" or "I don't know".
- *Not(X)*: The observable predicate *Not(X)* was used to make statements that an ontology class is no longer applicable to the situation. For example, a class may no longer be relevant since the relative may have denied questions about that class.

[5] Used Python package *Simple Transformers*: https://simpletransformers.ai/ (accessed: 2023-06-29).

– *Question(X)*: The predicate *Question(X)* is the only hidden predicate in the presented MLN. It describes ontology classes that might still apply to the situation and that the chatbot should ask about. For example, the predicate *Question(Communication Problems)* describes that the ontology class *Communication Problems* might be relevant to the relative and the chatbot could ask a related diagnostic question.

In order to include medical knowledge for the calculation of a probability distribution, logical formulas were extracted based on the relations between individual classes defined in the ontology. With the help of the MLN, it is possible to use medical knowledge from the ontology and guide the user through diagnostic questions suitable for the conversational situation.

For this purpose, various relations have been used in the ontology to define the domain knowledge and extract logical formulas in conjunctive normal form that could be used in the MLN:

– Hierarchy relation: The hierarchy of the ontology was used because parent classes often only roughly reflect the subject matter. Therefore, in many cases, it makes sense to generate deeper knowledge and further specify the problem with additional diagnostic questions according to child classes.
– Triggers: Describes from which causes the agitation of the PwD can be triggered.
– Leads to: Describes which agitation leads to which further consequences.
– Exhibits: This is the relation between concepts of the superclasses *Person with Dementia* and *Agitation*. The relation therefore describes which agitation the PwD can exhibit.

The following equations represent the hierarchy (see Eq. 3), triggers (see Eq. 4), leads-to (see Eq. 5) and exhibits (see Eq. 6) relations in the MLN:

$$\begin{aligned} &\forall c, p \in \mathbf{\Omega}, p \in \mathit{parent}(c) : \neg \mathit{Know}(p) \vee \mathit{Question}(c) \\ &\forall c, p \in \mathbf{\Omega}, p \in \mathit{parent}(c) : \neg \mathit{Maybe}(p) \vee \mathit{Question}(c) \end{aligned} \tag{3}$$

$$\begin{aligned} &\forall a, b \in \mathbf{\Omega}, b \in \mathit{triggers}(a) : \neg \mathit{Know}(a) \vee \mathit{Question}(b) \\ &\forall a, b \in \mathbf{\Omega}, b \in \mathit{triggers}(a) : \neg \mathit{Maybe}(a) \vee \mathit{Question}(b) \end{aligned} \tag{4}$$

$$\begin{aligned} &\forall a, b \in \mathbf{\Omega}, b \in \mathit{leads\ to}(a) : \neg \mathit{Know}(a) \vee \mathit{Question}(b) \\ &\forall a, b \in \mathbf{\Omega}, b \in \mathit{leads\ to}(a) : \neg \mathit{Maybe}(a) \vee \mathit{Question}(b) \end{aligned} \tag{5}$$

$$\begin{aligned} &\forall a, b \in \mathbf{\Omega}, b \in \mathit{exhibits}(a) : \neg \mathit{Know}(a) \vee \mathit{Question}(b) \\ &\forall a, b \in \mathbf{\Omega}, b \in \mathit{exhibits}(a) : \neg \mathit{Maybe}(a) \vee \mathit{Question}(b) \end{aligned} \tag{6}$$

Assuming the chatbot would recognize a physical-aggressive type of agitation in the relative's chat messages. Since, for example, the relation *physical-aggressive behavior leads to less social contact with other people* is mapped in the ontology,

the chatbot would then ask further questions about the social behavior of the person with dementia:

$$\begin{aligned} &\textit{Less social contacts} \in \textit{leads to(Physical aggressive)} : \\ &\neg \textit{Know(Physical aggressive)} \vee \textit{Question(Less social contacts)} \end{aligned} \tag{7}$$

This enables the chatbot to output help texts on this topic as the conversation progresses. For each extracted relation, a formula with a certain weight $\omega \in \mathbb{R}$ is inserted into the MLN to model the strength of dependencies between predicates. Hierarchical relations were given more weight than other relations because it was assumed that it is more important to generate more information about the current class and thus to explore deeper into the ontology tree.

However, since the predicates are intended to model different degrees of knowledge, the formulas over the *Know* relation were assigned double weight. To model some desired constraints of the real world, we added several hard constraints to our MLN:

– Diagnosis questions may not be asked twice:

$$\forall c \in \mathbf{\Omega} : \neg \textit{Know}(c) \vee \neg \textit{Question}(c) \tag{8}$$

$$\forall c \in \mathbf{\Omega} : \neg \textit{Maybe}(c) \vee \neg \textit{Question}(c) \tag{9}$$

– Classes that are not applicable based on our knowledge may not be asked for:

$$\forall c \in \mathbf{\Omega} : \neg \textit{Not}(c) \vee \neg \textit{Question}(c) \tag{10}$$

– Ask only one thing at a time:

$$\forall a, b \in \mathbf{\Omega} : \neg \textit{Question}(a) \vee \neg \textit{Question}(b) \tag{11}$$

We computed the inference of our defined MLN using the RockIt software[6] from the University of Mannheim. With that, we were able to calculate a probability distribution that delivers a probability for each ontology class of whether the class is applicable to the situation of the relative. In predicting the best diagnostic question, we only consider ontology classes whose probability is above a chosen threshold $\tau \in [0, 1]$. The two ontology classes with the highest probability were then returned to the user so that the user can decide among different options for the next topic.

4 Results

The evaluation was performed on two different data sets. On the one hand, the expert-generated data was divided into five different folds in a k-fold cross-validation. As a result, five different models were trained using four folds respectively as the training set. On the other hand, within a small study, a total of ten relatives of dementia patients were instructed to describe to the chatbot that the PwD refuses to be washed (*Resisting care*). Thus, a total of 20 samples of real chat messages were collected, which were used for the second evaluation. In the process, the model was trained with the entire data set generated by experts.

[6] http://executor.informatik.uni-mannheim.de/systems/rockit/ (accessed: 2023-09-08).

4.1 Data Generated by Experts

For each class, the *true positives (TP)*, *true negatives (TN)*, *false positives (FP)*, and *false negatives (FN)* were evaluated. Due to the small number of data for the labels *Yes*, *No* and *Unsure* individual folds could not be evaluated for the corresponding classes. Furthermore, the metrics *Accuracy*, *Precision*, *Recall*, *Specificity* and *F1 Score* were evaluated for each class, with the average calculated across each fold (see Table 2).

Table 2. The table shows the calculated metrics *Accuracy*, *Precision*, *Recall*, *F1 Score* and *Specificity* for all classes: *Physical Aggressive (PA)*, *Physical non-aggressive (PNA)*, *Verbal aggressive (VA)*, *Verbal non-aggressive (VNA)*, *Resisting care (RC)*, *Local help*, *Long version*, *Short version*, *Yes*, *No*, *Unsure* and *None*. The average of all five folds was calculated.

	n	TP	TN	FP	FN	Accuracy	Precision	Recall	F1	Specificity
PA	11.25	8.8	68.6	3.2	2.4	0.93	0.73	0.79	0.76	0.96
PNA	14	9.2	64.6	4.4	4.8	0.89	0.68	0.66	0.67	0.94
VA	11.4	9	68.8	2.8	2.4	0.94	0.76	0.79	0.78	0.96
VNA	10.2	7.3	70.8	2	3	0.94	0.78	0.71	0.74	0.97
Resisting care	10.8	9.8	69.2	3	1	0.94	0.77	0.91	0.83	0.96
Local help	5.4	5	77.6	0	0.4	0.99	1.0	0.93	0.96	1.0
Long version	4.2	3.6	78.6	0.4	0.6	0.99	0.9	0.86	0.88	0.99
Short version	3	2.8	79.2	0.8	0.2	0.99	0.78	0.93	0.85	0.99
Yes	1.75	1.5	80.75	0.25	0.25	0.99	0.86	0.86	0.86	0.99
No	1.75	1.5	80.25	0	0.25	0.99	1.0	0.86	0.92	1.0
Unsure	2	2	81.75	0	0	1.0	1.0	1.0	1.0	1.0
None	19	18.8	64	0	0.2	0.99	1.0	0.99	0.99	1.0

Chat messages for the agitations *Physical aggressive* (F1 value of 0.76), *Verbal aggressive*, *Verbal non-aggressive* and *Resisting care* were recognized similarly well by the trained models. The best values were obtained for the *Resisting care* label (F1 value of 0.82), since with a recall value of 0.92 almost all messages related to this type of agitation were correctly classified. However, the highest precision value (0.83) was achieved for messages concerning the agitation *Verbal aggressive*. Just messages related to the agitation *Physical non-aggressive* were recognized less well than the other labels.

For the evaluation of the entire model, the *macro average*, *micro average* and *weighted average* were calculated for each metric. The macro average calculates the arithmetic mean of the corresponding metric across all classes. With the micro average, on the other hand, the corresponding metric is calculated using the sums of the TP, TN, FP and FN counts. Finally, the weighted average also takes into account the frequency of occurrence of each class when calculating the arithmetic mean. These metrics were evaluated once for all classes, but also evaluated only for these corresponding agitation classes due to the focus on agitation

detection. The results of these metrics are shown in Table 3. The evaluation of the model for all classes is about 6–11% better than considering the agitation classes alone. Thereby, the macro, micro and weighted average value for the F1 score is 0.75, 0.76 and 0.76, respectively.

4.2 Data from Relatives of PwD

In the second evaluation with real chat messages from relatives, the model recognized 100% of the descriptions correctly as an agitation of the *Resisting care* type (cf. Table 4). Furthermore, during the evaluation, there were three other messages describing other forms of agitation, which were also recognized without any error.

However, the chatbot also incorrectly recognized some messages of the *Resisting care* class as the agitation *physical non-aggressive* and thus achieved a noticeably worse precision value (0.375) for this class (with only 3 samples). However, because the experiment was designed primarily for the *Resisting care* agitation type and only three samples exist, the metrics for this class offer lower reliability.

5 Discussion

Below, we discuss some advantages of our system compared to the currently popular *ChatGPT*.

The core difference in the developed chatbot lies in the ontology as its main knowledge base for a domain-specific use case. While *ChatGPT*'s main purpose is to provide a conversation partner rather than validated information, its core data set is mostly raw natural language data itself, functioning as its knowledge. With that, *ChatGPT* is very well able to generate a fitting piece of dialogue data that is likely to be a suitable answer to the user's question, but shows limitations in areas where strong concept understanding and association between multiple concepts are required [3,20].

In contrast to that, the *eDEM-CONNECT* pipeline provides a solution in which its knowledge base (the ontology) comes in a form that is on the one hand human-readable and on the other hand provides a conceptional knowledge base for the chatbot. This approach, even though it is more elaborate in its development, allows theoretically faster domain learning for the machine and a result that can be better controlled by the domain experts. Especially in the healthcare domain, where wrong information can potentially cause great harm, it is important to validate such information by domain experts. Currently, our approach has only been tested on a small concept set within the ontology due to the lack of training data for the chatbot. Because of that, a direct comparison between *ChatGPT* and our bot is currently not possible. So any predictions on a larger data set with training data are speculative. Larger data sets provide more complexity and make correct classification of concepts more difficult. The possibility of false classification increases with the growing number of concepts and discrimination between them becomes harder when several concepts are

Table 3. The table shows the metrics for evaluating model performance for all classes using micro, macro and weighted average. The average was calculated over all five folds.

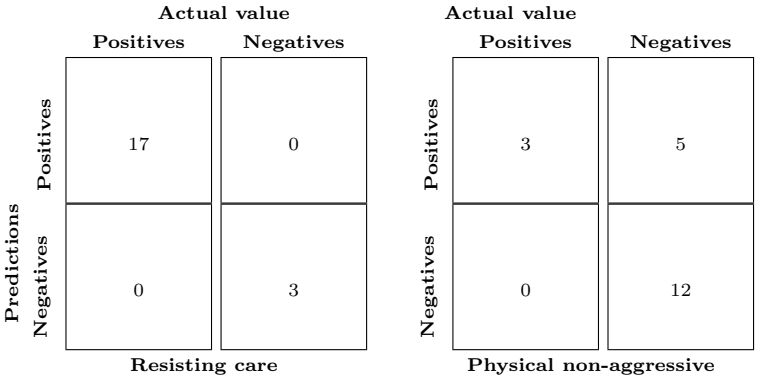
Metric	All classes	Agitation classes only
Macro Average Accuracy	0.97	0.93
Macro Average Precision	0.87	0.76
Macro Average Recall	0.86	0.77
Macro Average F1 Score	0.86	0.75
Macro Average Specificity	0.98	0.96
Micro Average Accuracy	0.97	0.93
Micro Average Precision	0.83	0.74
Micro Average Recall	0.84	0.77
Micro Average F1 Score	0.83	0.76
Micro Average Specificity	0.98	0.96
Weighted Average Accuracy	0.95	0.93
Weighted Average Precision	0.87	0.76
Weighted Average Recall	0.84	0.77
Weighted Average F1 Score	0.83	0.76
Weighted Average Specificity	0.97	0.96

more similar to each other. However, assuming the fact that there is a larger set of training data available in the future, the *eDEM-CONNECT* chatbot might be able to become the initially envisioned intelligent dialog assistant for family caregivers of people with dementia.

Furthermore, the evaluation has shown that the concept of detecting and classifying agitation descriptions in chat messages with a transformer model works fundamentally. The detection of the non-agitation classes (*Yes*, *No*, *Unsure*, *Local help*, *Long version*, etc.) also works exceptionally well, even with only limited data. The good results of the *Resisting care* class, could be due to the fact that *Resisting care* as a subclass of *Physical non-aggressive* covers a more narrow field than the other classes. However, evaluation of the real data set showed that some samples of the *Resisting care* class were incorrectly classified as *Physical non-aggressive*, too. The reason for this could be the definition of the hierarchy of these classes within the ontology and that the class *Resisting care* was modeled as a subclass of the class *Physical non-aggressive*. A restructuring of the labels could possibly improve this problem.

Responsible for some misclassification of the remaining classes could be the low degree of discrimination between agitations from type *Verbal aggressive* and *Verbal non-aggressive* or from type *Physical aggressive* and *Physical non-aggressive*, because these agitations were often described with very similar words in chat messages. In addition, samples were only labeled by one expert at a time.

Table 4. Confusion matrices of the two agitations from real chat messages of relatives of PwD: *Resisting care* (Left) and *Physical non-aggressive* (Right). The True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN) were calculated for each class. Furthermore, the table shows the calculated metrics: Accuracy, Precision, Recall, F1 Score and Specificity.



	Accuracy	Precision	Recall	F1 Score	Specificity
Resisting care	1.0	1.0	1.0	1.0	1.0
Physical non-aggressive	0.75	0.375	1.0	0.55	0.71

Due to the fact that humans label samples differently, this might increase the problem of low discriminatory power. An inter-observer reliability analysis could help to solve this problem.

An aspect that will still come up as a challenge in the future will be the transitioning condition between the anamnese step and the phase where the chatbot provides a solution to the user. In contrast to *Ada Health*, where the user is confronted with a large questionnaire, the transition here was intended to be more flexible and faster due to the users needing fast and pithy instructions. A shorter anamnese step will inevitably lead to more faulty instructions in the second phase. Therefore, there has to be some possibility to balance these two steps and maybe even go back to the anamnese status while the intervention phase has been reached.

Several approaches are conceivable here. An approach that has not been tested further would be to train a dialog system by using reinforcement learning and to give the system feedback in the case of good or bad predictions. A bad prediction therefore would be an unsuitable, further topic.

6 Conclusion and Further Work

In the *eDEM-CONNECT* project, we were able to develop a chatbot that uses a domain-specific ontology on challenging behavior of PwD as its main knowledge base. The chatbot is able to identify the concepts of a smaller subset of the

ontology when interacting with user inputs and can lead through a dialog in which it provides helpful information regarding a given situation with a PwD.

In the future, a project with a larger set of training data, enabling the chatbot to further deepen its knowledge of the ontology concepts, has very high potential. To realize the initial vision of the chatbot, a scaled-up program with complete ontology coverage and a clearer dialog process would be a suitable option. An alternative to that would be a bot without the ontology, which instead uses a domain-specific GPT, which are currently on the rise [20].

In both cases, a larger-scale user evaluation of the transformer model and the usability of the chatbot in general is also imaginable. Furthermore, the prediction of the next diagnosis question with the Markov logic network has not yet been reliably tested on real user data. The prediction of the next diagnosis question is based on the relationships between the individual ontology classes, which in some cases are difficult to prove medically. Therefore, other methods for predicting the next diagnostic question should be tested for further development to address this problem (e.g., prediction with reinforcement learning would also be conceivable).

Overall, the project provides a proof of concept for a domain-specific chatbot solution that incorporates an integrated ontology as its main knowledge base. This delivers a solid cornerstone for future projects that challenge domain-specific scenarios with the help of smart chatbots. In case of a successful future project that offers broader ontology coverage with good concept recognition, the chatbot will be better able to fulfill its initial project vision. This will improve the stability of the relationships of PwD and their caregiving relatives and professional caregivers will also profit from it. The care domain is continuously challenged by staff shortage and the bot can facilitate the situation by providing individual information to the users. With that, the movement to a retirement home could be prevented or at least delayed, relieving the current care situation in Germany.

References

1. Abdul-Kader, S.A., Woods, D.J.: Survey on chatbot design techniques in speech conversation systems. Int. J. Adv. Comput. Sci. Appl. **6**(7) (2015). https://doi.org/10.14569/IJACSA.2015.060712
2. Al-Zubaide, H., Issa, A.A.: OntBot: ontology based chatbot. In: Fourth IEEE International Symposium on Innovation in Information & Communication Technology, vol. 4, pp. 7–12. IEEE, Piscataway (2011). https://doi.org/10.1109/ISIICT.2011.6149594
3. Azaria, A.: ChatGPT usage and limitations. Preprint (2022). https://doi.org/10.13140/RG.2.2.26616.11526
4. Avila, C., et al.: MediBot: an ontology based chatbot for Portuguese speakers drug's users. In: 21st International Conference on Enterprise Information Systems ICEIS. ICEIS (Setúbal), vol. 21, pp. 25–36. SciTePress, Setúbal (2019). https://doi.org/10.5220/0007656400250036
5. Barros, A., Rajan, R.S., Nili, A.: Scaling up chatbots for corporate service delivery systems. Commun. ACM **64**(8), 88–97 (2021). https://doi.org/10.1145/3446912

6. Caros, M., Garolera, M., Radeva, P., Giro-i Nieto, X.: Automatic reminiscence therapy for dementia. In: Gurrin, C. (ed.) Proceedings of the 2020 International Conference on Multimedia Retrieval, pp. 383–387. ACM Digital Library, Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3372278.3391927
7. Chow, A.R.: How ChatGPT managed to grow faster than TikTok or Instagram (2023). https://time.com/6253615/chatgpt-fastest-growing. Accessed 24 July 2023
8. Clyburn, L.D., Stones, M.J., Hadjistavropoulos, T., Tuokko, H.: Predicting caregiver burden and depression in Alzheimer's disease. J. Gerontol.: Ser. B **55**(1), S2–13 (2000). https://doi.org/10.1093/geronb/55.1.S2
9. Deutsche Alzheimer Gesellschaft e.V.: Zum bundesweiten Tag der pflegenden Angehörigen: Angehörige von Menschen mit Demenz brauchen Entlastung - auch von Bürokratie (2023). https://www.deutsche-alzheimer.de/artikel/zum-bundesweiten-tag-der-pflegenden-angehoerigen-angehoerige-von-menschen-mit-demenz-brauchen-entlastung-auch-von-buerokratie. Accessed 07 Nov 2023
10. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 4171–4186. Association for Computational Linguistics, Stroudsburg (2019). https://doi.org/10.18653/v1/n19-1423
11. Feng, S.Y., et al.: A survey of data augmentation approaches for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 968–988. Association for Computational Linguistics, Stroudsburg (2021). https://doi.org/10.18653/v1/2021.findings-acl.84, https://aclanthology.org/2021.findings-acl.84
12. Gruber, T.R.: A translation approach to portable ontology specifications. Knowl. Acquisit. **5**(2), 199–220 (1993). https://doi.org/10.1006/knac.1993.1008, https://www.sciencedirect.com/science/article/pii/S1042814383710083
13. Kelle, N., Ehrlich, U.: Situation unterstützender und pflegender Angehöriger von Menschen mit Demenz. dza aktuell - Deutscher Alterssurvey (4) (2022)
14. Kothari, A.N.: ChatGPT, large language models, and generative AI as future augments of surgical cancer care. Ann. Surg. Oncol. **30**(6), 3174–3176 (2023). https://doi.org/10.1245/s10434-023-13442-2
15. Le Xin, T., Arshad, A., Salam, Z.A.B.A.: AlzBot- mobile app chatbot for Alzheimer's patient to be active with their minds. In: 2021 14th International Conference on Developments in eSystems Engineering (DeSE), pp. 124–129. IEEE, Piscataway (2021). https://doi.org/10.1109/DeSE54285.2021.9719410
16. Michalowsky, B., Kaczynski, A., Hoffmann, W.: Ökonomische und gesellschaftliche Herausforderungen der Demenz in Deutschland - Eine Metaanalyse. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz **62**(8), 981–992 (2019). https://doi.org/10.1007/s00103-019-02985-z
17. Müller, C., Paluch, R., Hasanat, A.A.: Care: a chatbot for dementia care. Mensch und Computer 2022 - Workshopband (2022). https://doi.org/10.18420/MUC2022-MCI-SRC-442
18. Ornstein, K.A., Gaugler, J.E., Devanand, D.P., Scarmeas, N., Zhu, C.W., Stern, Y.: Are there sensitive time periods for dementia caregivers? The occurrence of behavioral and psychological symptoms in the early stages of dementia. Int. Psychogeriatr. **25**(9), 1453–1462 (2013). https://doi.org/10.1017/S1041610213000768
19. Pinkert, C., et al.: Social inclusion of people with dementia - an integrative review of theoretical frameworks, methods and findings in empirical studies. Ageing Soc. **41**(4), 773–793 (2021). https://doi.org/10.1017/S0144686X19001338

20. Ray, P.P.: ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys. Syst. **3**, 121–154 (2023). https://doi.org/10.1016/j.iotcps.2023.04.003, https://www.sciencedirect.com/science/article/pii/S266734522300024X
21. Richardson, M., Domingos, P.: Markov logic networks. Mach. Learn. **62**(1–2), 107–136 (2006). https://doi.org/10.1007/s10994-006-5833-1
22. Ruggiano, N., et al.: Chatbots to support people with dementia and their caregivers: systematic review of functions and quality. J. Med. Internet Res. **23**(6), e25006 (2021). https://doi.org/10.2196/25006
23. Schirra-Weirich, L., Wiegelmann, H.: Typenbildung als Beitrag zur Weiterentwicklung von Versorgungsstrukturen für Menschen mit Demenz und ihren versorgenden Angehörigen. Ergebnisse einer Tandem-Studie im Rahmen des Modellprojekts "DemenzNetz StädteRegion Aachen". In: Schäfer-Walkmann, S., Traub, F. (eds) Evolution durch Vernetzung. Edition Centaurus - Perspektiven Sozialer Arbeit in Theorie und Praxis, pp. 59–76. Springer, Wiesbaden (2016). https://doi.org/10.1007/978-3-658-14809-6_4
24. Sosnowski, T., Abuazizeh, M., Kirste, T., Yordanova, K.: Development of a conversational agent for tutoring nursing students to interact with patients. In: Frasson, C., Mylonas, P., Troussas, C. (eds.) ITS 2023. LNCS, vol. 13891, pp. 171–182. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-32883-1_15
25. Sosnowski, T., Yordanova, K.: A probabilistic conversational agent for intelligent tutoring systems. In: Proceedings of the 13th ACM International Conference on Pervasive Technologies Related to Assistive Environments. PETRA 2020. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3389189.3397978
26. Subhalakshmi, Y., Shivani G.S., Sri Shandhya Devi, T., Sri Raksha Avanthiga, S., Ahila, R.: Androz chatbot for Alzheimer's patients. Int. J. Res. Appl. Sci. Eng. Technol. **11**(5), 3249–3256 (2023). https://doi.org/10.22214/ijraset.2023.52339
27. Thyrian, J.R., et al.: Burden of behavioral and psychiatric symptoms in people screened positive for dementia in primary care: results of the Delphi-study. J. Alzheimer's Dis.: JAD **46** (2015). https://doi.org/10.3233/JAD-143114
28. Valtolina, S., Hu, L.: Charlie: a chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In: CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter, pp. 1–5. ACM Digital Library, Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3464385.3464726
29. Varshini, M.P., Surabhi, S., Keerthan Kumar, T.G.: The companion chatbot for dementia patients. Int. J. Adv. Sci. Technol. **29**, 6582–6592 (2020)
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008. Curran Associates, Inc., Red Hook (2017)