DR 3.2: First prototype of user model construction algorithms featuring input from sentiment and interaction mining :

Maya Sappelli, Antoine Cully, Yiannis Demiris

TNO, Data Science, The Hague
Personal Robotics Laboratory, Imperial College, London
⟨maya.sappelli@tno.nl
a.cully@imperial.ac.uk
y.demiris@imperial.ac.uk⟩

Project, project Id:	EU H2020 PAL / PHC-643783
Project start date:	March 1 2015 (48 months)
Due date of deliverable:	February 29, 2017
Actual submission date:	February 29, 2017
Lead partner:	TNO
Revision:	final
Dissemination level:	PU

Executive Summary This document describes the first prototype of the PAL user modeling algorithm based on sentiment and interaction developed during the second year of the project in work package 3.

The overall objective of workpackage 3 is to adapt the behavior of the PAL system to each of its users. This adaptation is necessary to ensure 1) engagement of the user and 2) increased effectiveness in personal goal achievement. People tend to adapt their interaction style in a conversation based on observations such as the perceived sentiment of the conversation partner. For this purpose it is worthwhile to include sentiment in the interaction process, such that the robot or its avatar can adapt its behavior accordingly. Sentiment extracted from the child's diaries can assist in estimating the current well-being and emotional state of the child. Moreover, sentiment can be used as feedback mechanism to fine tune the user model.

This document first presents the state of the art in sentiment mining. One challenge that is specific for the application of sentiment mining in the PAL project is that sentiment mining is not typically applied to children's data. The language of children is still in development and is not reflected in existing trained sentiment detection models.

Since there is little data from the target group available that is already labeled for sentiment, data from the experiments have been annotated in order to improve existing sentiment mining algorithms. The data from the experiments is annotated with the perceived sentiment of the children. Analysis of the annotation reveals that the children convey negative sentiment in only 15% of the cases, thus they tend to be neutral or positive in their sentiment.

With the annotated data, the performance of an off-the-shelf sentiment mining solution is compared to an adapted approach which is adapted to accommodate children's language. The algorithm can detect the child's sentiment correctly in 82.5% compared to 76.9% of the off-the-shelf solutions. Further improvements that can be made to the algorithm are still under investigation.

In addition to sentiment information, the different interactions of the users with the PAL system provide important information about their knowledge. For instance in the context of the PAL system, the different responses of the users to quiz questions reflect their current knowledge about the disease and its treatment. This information is crucial as it may allow the system to personalize the educational path for each user according to his current knowledge level.

In this context, one of the main challenges is the amount of time required to collect enough data to infer the knowledge level of the user. In order to allow the system to rapidly adapt to the specificities of the user, the knowledge level inference needs to be achieved as quickly as possible. The difficulty relies in the fact that while quiz questions have the potential to provide information about the knowledge of the user on specific topic, the

actual information contained in one quiz response only informs the system that the user managed to correctly answer, or not, to the question. In order to capture the level of knowledge of one user on a specific topic, the system needs to ask several dozens of questions. Moreover, when we need to predict the knowledge level on several topics, the amount of interactions required to make an accurate prediction becomes too large to expect a rapid adaptation and personalization of the system.

In order to address this problem, we introduce in this document an interaction mining approach that leverages the information collected from all the users of the system to make rapid and accurate estimates of the knowledge level of a new user thanks to knowledge transfer. In particular, this approach enables the PAL System to require up to 10 times fewer interactions than traditional approaches, to reach the same level of accuracy.

1 Role of sentiment and interaction mining in PAL

WP3 focuses on the personalization and the adaptation of the PAL system to the preferences of and and specificities of each user in order to ensure their engagement and adoption of the PAL system. In particular the work package aims to personalize the interaction of the user with the system to optimize the child's learning process. Sentiment and interaction mining are important aspects in this regard as it can guide action selection. Sentiment is an element of the current state of the user. Responding to this state is likely to improve bonding between PAL and a child [6]. In addition to its inherent link to WP3, sentiment mining is related to the PAL ontology (WP1), and can provide feedback on the child's progress and state which is relevant for WP2. Moreover it will receive input from WP4 and the analyzed sentiment of the input could be used for real time feedback (WP5). Currently the sentiment mining algorithm is generic to all users. Personalization of sentiment detection could be an improvement, in which case the algorithm can learn from on-line behavior and feedback. An example is using the feedback from the AffectButton [4] directly to improve (personalized) training of the sentiment analysis module.

Exploiting the information contained in the interactions with the serious games of the system may provide precious information about the knowledge level of the users on the different topics covered by the PAL System. This information can then be used to personalize the selection of the topics of the quiz questions to offer to the children questions that are not too difficult nor too easy, but just with the right level of difficulty. Such topic selection approach allows the user to remain in his "zone of proximal development", which is known to provide optimal educational path[29].

2 Tasks, objectives, results

2.1 Planned work

The objectives of this deliverable in the Work Package 3 were to provide a first prototype of sentiment and interaction mining. To reach this objective, our goals have been:

- To review the current state of the art in sentiment mining in order to see whether there are off-the-shelve solutions that could be used for the sentiment mining goals in the PAL Project. This is described in Section 3.1
- Analysis of the data from the diaries collected during the first year experiments and camps to see which sentiments are expressed in the PAL context by the children, including the annotation and preparation of training data for sentiment mining algorithms.
- Implementation of a first prototype of the sentiment mining which can detect the sentiment of new sentences and which can be integrated with the PAL system.
- Evaluation of the effectiveness of the trained algorithm before it is integrated with the PAL system
- To review the state of the art in interaction mining and cold start problem
- To design a user model for knowledge level that can adapt to changes over time and directly use raw (binary) data.
- To incorporate in our user model a collaborating filtering approach to allow the model to make predictions based on sparse data
- To evaluate the accuracy of our user model based on real data collected during the experiments that took place last year.

The following section will detail the work and the results that have been achieved during the second year of the project.

3 Actual work performed

3.1 Review of the state-of-the-art

Below we briefly discuss the literature in sentiment mining. In Natural Language Processing a distinction is made between the analysis of sentiment (polarity classification): positive, neutral and negative, and emotion analysis: anger, sad, etc. In the PAL project we have focused on sentiment mining

so far, since most off-the-shelf solutions offer this type of analysis. Moreover, emotion detection is more complex [1]. Only if the emotions expressed by children to the robot or avatar show a sufficient amount of variation and occurrence to make a more fine-grained detection relevant will we consider an extension to emotion detection.

Most sentiment mining research is directed at the analysis of (online) reviews, blogs, forum texts, and tweets [3][15][19]. The purpose is usually to provide a grip on customer experience [19], for example on how a company or a product is perceived. In the PAL project, sentiment mining is about how a child is feeling - for example how his or her day was, or how he/she felt about glucose levels.

Current sentiment algorithms often use a lexical resource such as *SentiWordNet*. This is a dictionary in which words are labeled with a positivity, negativity, and objectivity score. These scores are obtained using semi-supervised methods [2]. Assessing the sentiment of a text then consists of determining the complete score for that text.

It is important to take boosters and negations into account. For example, 'not good' can be seen as negative sentiment. However 'good' is a word with positive sentiment, and only when the words are combined the sentiment is perceived as negative. Another example is 'ok!!!!!', ok in itself might have a neutral connotation. However, the addition of multiple exclamation marks suggests enthusiasm that might lead to a positive interpretation of the phrase.

There are two approaches to take into account boosters and negations. The first is a *symbolic* approach, where scored n-grams are combined with rules and thresholds. For example, if 'not' preceeds a word with a positive sentiment, the sentiment is reversed [18]. Moreover if the combined score of a text exceeds a certain threshold, the text is classified as 'positive', otherwise as negative. Another *symbolic* method is to use semantic similarity, which can be calculated by using *WordNet* [11]. This is a resource in which words are connected to each other based on their semantic relations, e.g. X is a synonym of Y. The length of a path from a word in the document to semantic words such as "good" and "bad" in *WordNet* can be used as an indicator of sentiment [5]

In the second approach, *machine learning* is applied to automatically learn the relations between labels or scores and the individual elements [3]. The advantage of such a machine learning approach is that it takes the context of the data into account: 'ok' in one context can be positive, while in another it can be negative. Deep learning is also successfully applied to sentiment mining [27]. Machine learning is often used for emotion detection [7, 25, 14, 10]. However, machine learning techniques typically require a large amount of (labeled) training data. Existing data sets cannot be used as they are not targeted at the language children express.

There are very few approached of sentiment mining directed at children's

language. An approach directed at child language is important, since their language use and sentiment interpretation might differ from that of adults. One example of analysis of language for children comes from a Thai sentiment resource which has been developed for Thai children stories [20] . It was created from a translation of English terms in SenticNet2. Using Support vector machines 75.67% accuracy is achieved in detecting the sentiment of the children stories. This is low compared to sentiment detection on reviews (typically ¿ 85% accuracy [28]).

Alm, Roth, and Sproat [1] also researched machine learning techniques for sentiment and emotion analysis of sentences from children stories. The authors concluded that assessing the emotion of a text (i.e. sad, angry) was a difficult task and inter annotator agreement was low. Therefore, their model aimed at classifying a sentence as either positive, negative, or neutral. A dictionary of positive and negative words was used for classification. Additionally they found that sequencing, taking into account the emotion of adjacent sentences, was beneficial in some cases, showing that taking into account emotions from related sentences can possibly improve classification results.

Concluding, Off the shelf solutions are not targeted at children's language or do not have Dutch and Italian models. There are two directions that can be taken. The first is to train a system from scratch. The second is to use a pre-trained system and adapt it to make it better suited for children's language. Considering the small amount of training data that is available, the latter approach seems more feasible

3.2 Analysis of the sentiment data provided by children

During the PAL project, there have been three Diabetes camps in which data has been collected on the sentiment of children in the age of 8 to 14 with Diabetes. At several occasions, children could provide a note or some text on how they felt.

In total there were 395 Dutch entries by 36 children. The median number of words in an entry was 8, with a minimum of 1 (17 entries) and a maximum of 42 words. Italian entries were not taken into account as these were predetermined sentences that were entered as a usability test.

All entries were annotated by three annotators. Inter annotator agreement, measured using Cohen's kappa was 0.85. For approximately 40% of the entries a ground truth provided by the children themselves was available in the form of happy, sad, neutral labels, which they selected using an emoticon. For approximately 30% of the entries a ground truth was available in the form of a number between 1 and 6, where 1-2 are interpreted as negative, 3-4 as neutral and 5-6 as positive. For the remaining 28.6% no ground truth was available. Table 1 depicts the distribution of positive, negative and neutral sentiments in the data. The ground truth data was not

collected with the purpose of evaluating the text entries, but rather as an independent source of sentiment expression. Because the elicitation method used with the children varied and did not match the elicitation method for the annotators, inter-annotator agreement between child and annotator can not be determined reliably.

Table 1: Percentage of positive, negative and neutral sentiments in the data

	ground truth	annotators
negative	7.6	15.2
neutral	20.3	42.3
positive	43.5	42.5
none	28.6	0

Table 1 shows that the data is skewed to the neutral and positive sentiments. Moreover, the children themselves seem to assess their own sentiment as more positive than the annotators. This can also be a bias created by the emoticons that were used to determine the label 'happy', 'sad' or 'neutral'. The children chose the happy emoticon in 70% of the cases, while they chose a happy sentiment using numbers only in 48% of the cases. The annotations by the external annotations are used for further evaluation in order to overcome these discrepanties.

In order to gain insight into which words are indicative of certain sentiments, a logistic regression model was trained. The top ten words from the model, based on their coefficient values per sentiment are displayed in Table 2 starting with the highest one on top. The results show that certain words such as sugar have an interpretation that is strongly related to the context of Diabetes. This indicates that it may be necessary to adapt a sentiment classifier not only to the language used by children, but also to the context of the sentiment expressions

3.3 Prototype of sentiment mining algorithm

In the first version of the algorithm we used Sentiwordnet to tag the terms. Since Sentiwordnet is an English resource, a translation from Italian or Dutch was made. This introduced some translation errors. Therefore in this approach we have used Pattern [24]. This is a pre-trained lexicon-based algorithm which provides a polarity and subjectivity score for each sentence. This is a numeric indication between -1 and +1 which indicates how positive or negative the expressed sentiment is. The Pattern algorithm is used because it offers a Dutch trained algorithm and its approach is similar to Sentiwornet. We compare the off-the-shelf Pattern results to an approach, Pchild, in which Pattern is extended by including boosters and expert ontology information. The Pchild sentiment mining algorithm consists of five steps:

positive	neutral	negative
fantastisch (fantastic)	echte (real)	uitgegleden (slipped)
blij (happy)	regent (raining)	jammer (pity)
ja (yes)	leuk (nice)	getypt (typed)
zin (inclination)	goed (good)	suiker (sugar)
of (or)	eigen (own)	bolesde (?)
voordat (before)	dag (day)	stom (stupid)
leuk (nice)	als (if)	moe (tired)
lekker (yummy)	229 (229)	helaas (unfortunately)
leuke (nice)	lekker (yummy)	duizelig (dizzy)
gezellig (cozy)	was (was)	baby (baby)

Table 2: Top ten words with highest sum of absolute coefficient values in a logistic regression model

- The first step is to extract booster punctuation such as exclamation marks, these will be used as a multiplier for the score of the word accompanied by the punctuation. Negation words such as 'niet' (not) are extracted as well and are used as polarity reversion mechanism.
- The second step is to parse the text into sentences and words. For this purpose Pattern is used. Pattern generates a full parse tree with part of speech tagging and lemmatization. Moreover, Pattern provides a polarity score for each word in the sentence and deals with multiword terms that include negation.
- The next step is a polarity score based on similarity to an expert ontology. This ontology describes several emotion related terms used by children and whether these are associated with a positive or with a negative sentiment. Examples are 'blij' (happy), 'trots' (proud), 'boos' (angry), 'saai' (boring). We have trained a Word2Vec model [17] on the Basilex Corpus[1] . The Basilex Corpus is a collection of text that stems from children's educational and fictional books. These texts are thus targeted specifically to children. The Word2Vec model creates vector representations of words based on the context in which a word occurs. This makes it possible to determine the distance of a term from the children's diary to the expert ontology. The classification (positive or negative) that is closest to the term at hand, determines the polarity score, 1 for positive, -1 for negative. Only terms that exceed a similarity score of 0.5 are considered.
- The scores from Pattern and Word2Vec are merged by selecting the maximum score. This generates a positive bias, as we observed in the

[1]parameters: size=200, min_count=10, window = 8

children as well. A next step is to train an ensemble method to find the optimal merge strategy.

- As a final step, the booster and negation adaptations are used to boost or reverse the polarity score. A classification into positive, neutral or negative is made using thresholds: > 0.2 for positive, < -0.2 for negative, and in between is neutral.

3.4 Evaluation of sentiment mining algorithm

We have used the data that was collected during the Diabetes Camps in the Netherlands. This data is described in section 3.2. In this section we describe the results with the off-the-shelf solution Pattern. We compare it to our own approach Pchild which is an adaptation to the original Pattern approach targeted at children's language.

		predicted		
		negative	neutral	positive
actual	negative	31	23	6
	neutral	27	130	9
	positive	6	20	143

Table 3: Confusion matrix of the Pattern method

In table 3 the confusion matrix of the Pattern method is presented. The Pattern method has an accuracy 76.9% (balanced 71.5%). In the confusion matrix we see that positive and negative labels are confused in 0.03% (12.1% of the errors), this is a more severe error than if neutral labels were confused. In table 4 the confusion matrix of the Pchild method is presented. Here we have reduced the confusion of positive and negative labels to 0.02% (11.6% of the errors). Both methods are not really accurate in determining true negative labels with Pattern having 51.6% accuracy and Pchild 38.3% accuracy for the negative labels). This is likely caused by context-dependent sentiment associations (such as a negative association with the word 'sugar') that are not reflected in the Pattern and Pchild method. Overall accuracy of the Pchild method is 82.5% (balanced 73.1%).

		predicted		
		negative	neutral	positive
actual	negative	23	30	7
	neutral	4	156	6
	positive	1	21	147

Table 4: Confusion matrix of the Pchild method

One example of words that have a different connotation in adult language and child language is the word 'naar'. This word can be interpreted as a preposition 'naar school' (to school), or as a sentiment 'ik voel me naar' (I feel bad). In child language this word is almost always as a preposition, whereas Pattern interprets sentences with this word as negative, because it does not user the parse type into account.

Using the word2vec method, words that tend to be used by children in a positive or negative context can be found. An example is 'echt' which tends to have a positive connotation, for example in the meaning of 'echt leuk' (really fun). In essence the word2vec method can be used as a data-driven method to find booster words or synonyms. Since the word2vec model is trained on children's language, these detected booster words or synonyms are specific to children.

3.5 Review of the state-of-the-art in interaction mining

Within the PAL System, children are invited to play quiz games that are designed to improve their knowledge about their disease and its treatment. For instance, through interactions with the quiz games, children can learn how to count carbohydrate in their meals, how to recognize hypoglycemia symptoms, or how to follow their diet. In total, there are 29 different topics covered by the quiz questions that more or less complex and designed for the different learning stages of the children.

The children's answers to these quiz questions could provide information on their knowledge level. This information is crucial for intelligent tutoring systems like PAL because it can be used to personalize the educational path provided to the children. In order to estimate the knowledge level of each child on the different topics covered by the quiz games of the PAL application, we introduce in this report a user model that is built based on information provided from interaction mining. To fulfill this goal, the user model needs to face two scientific challenges: 1) how to deal with sparse data and 2) how to adapt the predictions over time according to the progress of the users.

Optimizing the teaching sequence is a long-standing question in the domain of Intelligent Tutoring systems. Several works have been made in this direction while considering different educational goals, different constraints on the available data or different models [13]. For instance, they suppose that they have access to prior knowledge about the domain (relation between the different knowledge components or topics), or that they can rely on an extensive amount of data or rely on a pre-defined model for the progress of the student. Two models are used in the vast majority of the literature for tracking the progress or knowledge level of students: Bayesian Knowledge Tracing [8], which is based on a Bayesian model, and performance factor analysis [21], which is based on a logistic regression function. Both of these

approaches assume that the progress of the student will follow a particular model, which is governed by parameters that need to be defined to fit the observations that come either from the entire population of students or from one student in particular. In general, finding the values of the parameters is a challenging problem, first because of the lack of data and second because several sets of parameters can explain the data. Moreover, these models are making the assumption that every student follows the same developmental path (typically a logistic regression function), while it is expected that each child experiences several learning plateaus at different stages of their development[12]. While, some of the proposed approaches consider the potential progress of the user according to the number of interactions with the system (number of answered questions), none of them take into account the exact moment when these interactions took place, which may provide additional information. For instance, if the child does not play with the PAL system for an extensive period of time, it is likely that his/her knowledge level has significantly changed, because he/she has forgotten the notions previously learned or because he/she acquired some knowledge from external sources (like at a camp). For all these reason, the prototype of our user model, based on interaction mining, introduced in this deliverable is a *model-free* approach (i.e., the model does not suppose that the development of child follows a predefined function).

In the context of the quiz games in the PAL system, a large amount of interactions is required to make an accurate estimation of the knowledge level of the child on one particular topic. This comes from the fact that each question returns only a limited amount of knowledge: It only informs the system if the child managed to response correctly to the question or not. This binary information is not enough to infer the actual knowledge level on a topic and the system needs to accumulate several dozens of responses in order to make an accurate assessment that is not biased by non-knowledge related factors (e.g., random guess, mistakes caused by distraction or ambiguous formulation of questions). For instance, 10 data points only provide a rough approximation of the user level, while 100 data points provide a more accurate estimation. This difficulty is amplified by the number of topics on which we would like an estimate of the knowledge level. For instance, if twenty questions are used to make an (potentially inaccurate) estimate on one topic, then more than 580 questions are required to form a global estimate of the user's knowledge level. This large amount of questions most of the time represents a limitation for intelligent tutoring systems, as they may be unable to provide personalized educational path as long as the estimation of the knowledge level of the user is not completed. Designing a model that can account for sparse data and, therefore, provide accurate estimate with only a little amount of data is of crucial importance.

Providing accurate estimate of users' preferences given a limited amount of data or for new user is a well documented problem name "cold start".

For instance, this a common challenge for commercial recommendation systems, like Amazon or Netflix, as they need to suggest items or films to new users. In this domain, two main approaches have emerged: *Content-based* recommendations and *Collaborative* recommendation[23].

In the first case, the main concept consists in finding items that are similar to those that the user appreciated in the past. This is achieved by using similarity measures on the attributes of the items[16]. For instance, if the items A and B are similar, then it is likely that users who liked the item A will like the item B as well. The concept can be also centered around the users: if two users are similar (e.g., according to their age, origins, history), then they are likely to share the same preferences. Unfortunately, this kind of approach cannot be applied to our problem because the attributes of the different quiz topics contain only the name of the topic. On the other side, while the demographic information of the users contains a significant amount of information, the link between these attributes and level of knowledge is not consistent. For instance, the age of the users does not always reflect their level of knowledge, as the diagnostic of the disease may happen at different period and the social environment may be different.

The second approach, also named collaborative filtering, uses information collected from previous users (for example, how much they rated a particular item) to base its predictions for a user who never encountered this item, or for a new item never presented to the community [26]. For instance, if a group of users like both item A and B, it is likely that a new user who likes item A, likes the item B as well. This approach can be used in the context of the PAL system because we can infer that if several users have similar knowledge levels on some topics, then it is likely that their knowledge level is similar on the other topics.

In this deliverable, we introduce a user model based on interaction mining that is composed of a temporal, model-free predictor combined in which we use collaborative filtering methods to allow the model to make prediction even with a limited amount of data. The next two sections present these two main components of the user model.

3.6 Prototype of a temporal model-free knowledge level predictor

In this section, we introduce a knowledge level predictor which does not rely on a pre-defined model or function to explain the current state or progresses of the children (this type of predictor is called "model-free") and that adjusts its prediction over time to reflect the variations of the users' knowledge levels.

The method proposed in this deliverable is based on Gaussian Processes [22], which is a statistical model-free approach that can be used as a non-linear regression algorithm. It uses data collected in a continuous space (which can be a spatial and/or a temporal domain) and predicts the value

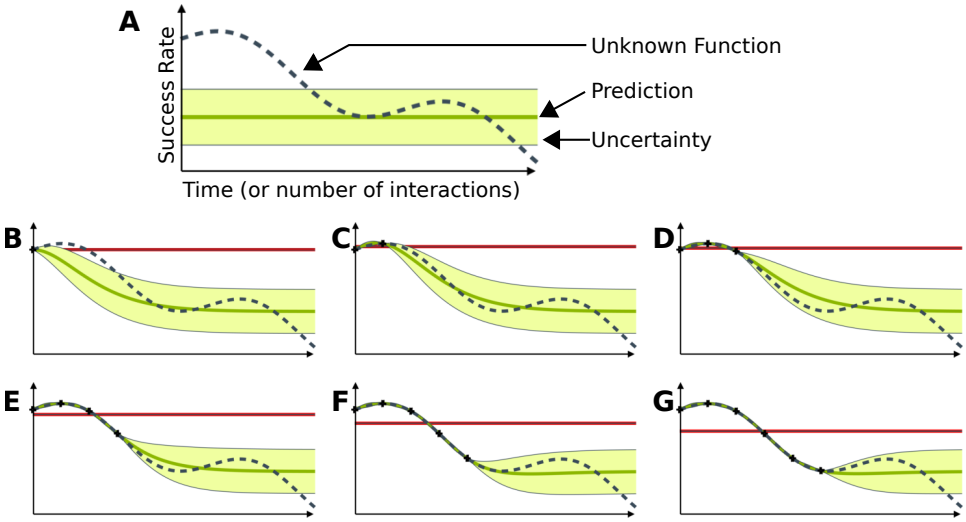

Figure 1: Temporal property of the proposed approach. The model predicts the success rate of the user on one topic (probability that the user will respond correctly to one question). The prediction (green line) corresponds to the mean of the predicted probability distribution, while the yellow area represents the uncertainty of the prediction. The unknown function (here a toy problem) that the model needs to approximate is depicted by a dashed line. For this illustration, we used an arbitrary function to highlight the ability of the proposed algorithm to adapt to personal development of the users. A) Initially, when no data has been recorded, the model outputs a constant success rate with a high uncertainty (large yellow band). B-G) Progressively, data is collected and the predictions are refined. The uncertainty is reduced in the area where data has been collected. The red lines indicate the average value of the recorded data. This highlights the importance of making predictions that depend on the acquisition time of the data.

of the underlying and unknown function at locations where no data has been collected yet. However, instead of predicting a single value (similar to most of the regression algorithms), Gaussian Processes predict the probability distribution of the possible values at each point of the domain. This statistical representation provides information about the confidence of the predictions. For instance, the probability distribution is narrow when the uncertainty on the prediction is low, and wider when the uncertainty is high.

In the proposed approach, we use the regression abilities of Gaussian Processes to take into account the moment at which each data point has been recorded in order adapt its prediction over time. Figure 1 illustrates the temporal property of proposed approach.

The second difficulty in the context of quiz games in the PAL system

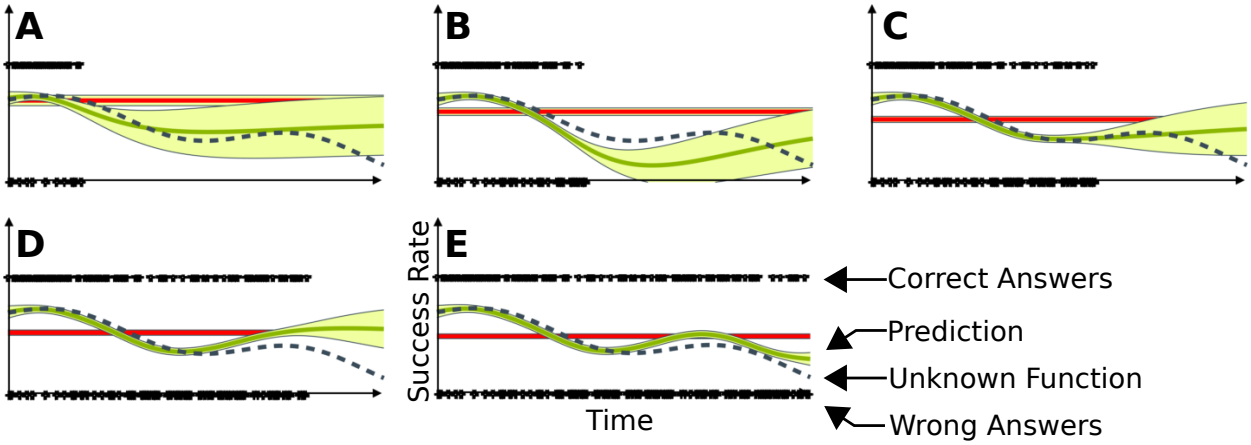

Figure 2: Success rate predictions based on raw interaction information (binary information). In this toy problem, the unknown function (dashed blue line) represents the success rate of the user on one topic. At each time step, the algorithm receive a new data point (positive or negative) which is generated according to the success rate defined by the unknown function. The objective of our algorithm is to predict the success rate of the user (i.e., approximate the unknown function) by only using the interaction information. A-E) Successive predictions provided by our algorithm. Each panel depicts the predictions after receiving 40 interaction data points. Similarly to the previous figure, the red line corresponds to the average of the data points.

is that the collected data does not represent the actual knowledge level or success rate of the user. The information provided by the system after each quiz question is only if the user managed to respond correctly to the question or not. For this reason, our approach needs to be able to deal with binary information in order to predict continuous success rate. To achieve this, we used a property of the Gaussian Process that allows the algorithm to average recent data points by considering that each of them is really uncertain. This property, which is rarely used in the literature, allows our algorithm to both change its predictions over time and use binary data. Figure 2 illustrates these properties.

In order to assess the level of the user on different topics, the same model can be replicated and specialized for each topic by only taking into account the questions related to the considered topic.

3.7 Prototype of collaborative filtering for knowledge level estimation using sparse data

In the previous section, we showed that our model is able to follow the changes of the user level over time while considering only binary information. However, another challenge in the context of the PAL system is the lack of

data. In order to achieve accurate predictions, a significant amount of data is required. For instance, in Figure 2, several dozens of data points are required to update the predictions. In practice, this means that the children have to answer to dozens of questions before that our user model collects enough data to make predictions. This also means that during this period, the user could not benefit from personalized behavior, which may hurt his experience with the system.

Therefore, the second objective of the user model is to minimize the amount of data required to make accurate predictions. To achieve this objective, we combined collaborative filtering techniques, detailed in section 3.5, with the user model described above. The main idea of our approach is to use the data collected from the previous users in order to make more accurate predictions quicker. This is achieved by using the data collected with the current user to determine which of the previous users has a model that explains the observed data best (also named the likelihood of the predictions). With this approach, we do not need to compare attributes of the users because it only relies on the collected data.

We combined this approach with our user model by substituting the "mean term" of the Gaussian Process with a fusion operator. The mean term is used to define the output of the model when there is no data. For example, in Figure 1.A, the mean term is defined as a constant value. However, an arbitrary function can be used as a mean term in a Gaussian Process [9, 22]. In our case, the fusion operator is a linear combination of the predictions coming from the previous users. However, more complex fusion operators can be considered in future works, like neural networks. The advantage of using a fusion operator in place of the mean term is that the mean term provides the global shape of the prediction while the rest of the Gaussian Process encodes for the local specificities of the new user. For instance, if the new user is very similar to a previous user on most of the topics, but very different on one of them, the fusion operator will suggest to use the prediction of the previous user as a first guess, while the Gaussian Process will progressively refine the predictions by taking into account the user's differences. The parameters of the fusion operator can be automatically determined by optimizing the "likelihood" of the model given the data[22]. For instance, in the case of a linear combination, the weights can be determined with this procedure. In this case, the algorithm will determine automatically the combination of the predictions coming from the previous users that best matches the observed data while being as simple as possible. The combination of the fusion operator and the likelihood optimization of its parameters provided by the Gaussian Process leads to an adaptive collaborative filtering method. Figure 3 illustrate the general structure of the user model used for interaction mining.

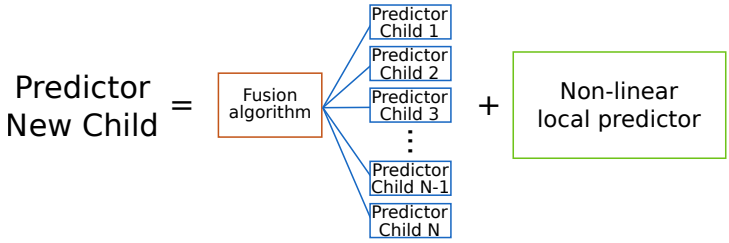

Figure 3: General structure of the user model used for interaction mining. The Non-linear local predictor corresponds to the Gaussian Process, while the fusion algorithm corresponds to the "mean term" used by the Gaussian Process.

3.8 Evaluation of the prototype of user model based on interaction mining

We evaluate our user model by using quiz data collected from children during the experiments that took place the first year of the project (at the hospital, home situation and camp). Among the children who participated to these experiments, we selected the 17 ones who answered more than 75 questions. With the collected data, we computed the success rate of each child on the different topics of the quiz game and we used those success rates to create virtual children with the same success rates. These virtual children respond to the quiz questions with the same success rate as the real children. For this evaluation, the goal of our model is to predict the actual success rate of the user (which is unknown by the model). At each time step, a question from a randomly select topic is asked from the virtual child and its response is recorded and used to update the prediction. Figure 4 shows the prediction accuracy of our model compared to 2 other approaches. This experiment has been replicated 17 times on 4 different contexts, in which the number of quiz topic varies. For each of the replication a different child is selected among the 17 ones and used as "the new user" while the 16 other children are used as "previous users" (this approach is often used as a cross validation technique). We considered 4 different cases where the number of topics increases in order to illustrate the scalablility of our approach.

We compared the results of our approach against two other techniques. For the first approach, we consider predictions using only the collected data from the new user. For the topic in which no data has been collected, the predicted success rate is 50%, otherwise the success rate is the number of correct answers divided by the number of question asked. For the second approach, we used the predictions that come from the closest previous user based on the predictions made by the first approach, explained just above. We named this second approach the "full-collaborative filtering" method.

The results show that our approach behaves initially as a "full-collaborative

filtering" approach, as the quality of the predictions is similar to the predictions made by the collaborative filtering approach. However, when the amount of collected data increases, the quality of the predictions made by our approach rapidly increases and becomes significantly better than the compared approaches (p-values$<$ 0.0014 after 300 interactions and 36 topics). At the end of the experiments (i.e., after 300 questions asked), the performance of the method based only on the user data become more accurate than the collaborative filtering one. However, in all the cases and replications made, our approach shows the best results compared to the other approaches. Moreover, the experimental results demonstrate that the performance of our approach is not impacted by the number of topics, while the other approaches (in particular the "user data only" one) require more interactions to make predictions of similar accuracy when the number of topics increases.

3.9 How dealt with review comments

R12 *The existing sentiment analysis provides the minimum of 3 behavioral classes. Favoring a more realistic approach it is recommended partners to extend the classes of behavioural status.*

In order to extend the sentiment mining with more dimensions, we have first looked at the data that we collected in the various experiments. The preliminary results extracted from the first year experiments reveal that the range of emotions expressed by children is not sophisticated enough to justify the use of more dimensions of sentiment analysis from texts. For example there were only few negative examples, suggesting that it is not worthwhile to distinguish emotions (anger, fear) further. This has resulted in the choice for an algorithm that determines a direction of sentiment (negative, neutral, positive) and a numeric polarity score that indicates how strong the sentiment was expressed.

R14 *Although the adopted mining tools for adaptive action selection can deal with sparse data, it is worthwhile the work of this WP to justify the impact of sparse critical data.*

In this deliverable, we highlighted the impact of sparse data on the accuracy of the user model predictions and how it is crucial to use algorithms that can leverage the use of data from other sources (in this case, other users) to make more accurate predictions. While the method introduced in this report only considers data related to the quiz games, we plan to extend this approach to other parts of the PAL System.

R13 *The evolving guidance strategy according to model drift for each individual patient can likely contribute for the enrichment of knowledge referring*

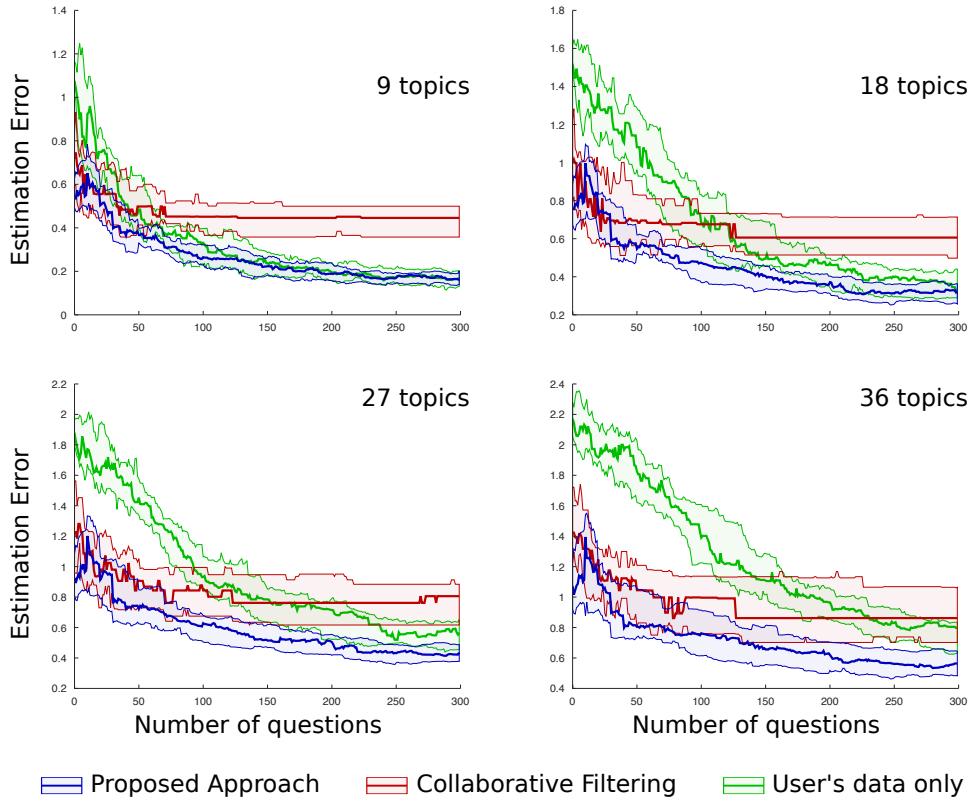

Figure 4: Evolution of the estimation error according to the number of interactions. The estimation error is computed as the euclidean distance between the prediction made by the different models and the actual success rates of the virtual child (the success rates form a vector a real value, between 0 and 1, where each dimension corresponds to a different topic). Four contexts have been considered in which the number of topics changes in order to highlight its impact on the performance of the compared approaches. For each context and each of the compared approach, the experiment has been replicated 17 times. Each of the replication considers the data recorded from a different (real) child.

to Personas Development. It is thus suggested Partners to explore this alternative.

Related to the answer above (R14), the meta information found by the approaches introduced in this document (like clusters of user, or general trends) are planed to be used to enrich the personas used in the project, when data will be recorded on long-term usage.

4 Conclusions

There are several aspects that make sentiment and interaction mining for the PAL project a challenge. First, it is a challenge to elicit sentiment data about personal feelings from children. They tend to give short answers only and find it difficult to express their feelings. Moreover, there simply is not that much variation in sentiment that is expressed, most expressions are neutral or (slightly) positive. Second, because of the intrinsic property of interaction data that defines the user's knowledge level (binary and temporal).

Moreover, the sparseness of data available for both the sentiment and interaction mining and the skewedness in the data restricted the options for sentiment analysis. Negative sentiments do not occur very often, but if they do occur they should not be missed.

Overall, the WP-tasks provided several major outcomes:

- An analysis of the state of the art in sentiment mining has revealed that there are no off the shelf solutions available. Solutions are typically not directed at children's language.
- An analysis of the sentiments expressed in writing by children in the context of the Diabetes Camps. It shows that the children do not express a wide range of emotions in text and that the expressed sentiments do not tend to be negative.
- An implementation of a first prototype of a sentiment classification algorithm.
- An offline evaluation of the developed prototype which shows that the adapted algorithm performs better than the existing sentiment detection methods.
- An analysis of the start of the art in recommendation system and models of user's knowledge level that highlight that the challenges that PAL need to face have never been specifically addressed in the literature.
- An implementation of a first prototype of user model for knowledge level estimation based on interaction mining.

- An implementation of a method to deal with the cold start problem in the user-model by using data from the other users of the system.
- An offline evaluation of the developed prototype based on real data that demonstrates that our approach for knowledge level estimation is at least an order of magnitude faster and more accurate than other approaches.

Currently, the sentiment mining algorithm is further improved. Improvements include a better Word2Vec model which is based not on what children read, but on what they write themselves (BasiScript). Moreover, machine learning algorithms, such as logistic regression and naive bayes are explored to reduce the dependence on language dependent sources such as Pattern and SentiWordNet. Finally, a stronger connection to the ontology developments in the other workpackages will be made, which will also allow for a better integration of the sentiment mining module into the PAL client and the experiments.

The knowledge level estimation approach will be integrated in the PAL system and is planned to be used in the next sessions of experiments. In particular, the estimate will be used to select topics on which the user has a success rate close to 50% which represents topics that are not too easy (the score on such topic is close to 100%) or too difficult (for multi-choice question with 4 answers, a random guess approach, a theoretical user will get 25%). This selection approach will allow the users to remain in his zone of proximal development, which is known to provide an optimal educational path. We will investigate if the impact of this method on the engagement of the users to the system.

References

[1] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.

[2] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[3] Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558, 2009.

[4] Joost Broekens and Willem-Paul Brinkman. Affectbutton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641–667, 2013.

[5] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[6] Franziska Burger, Joost Broekens, and Mark A Neerincx. A disclosure intimacy rating scale for child-agent interaction. In *International Conference on Intelligent Virtual Agents*, pages 392–396. Springer, 2016.

[7] Soumaya Chaffar and Diana Inkpen. Using a heterogeneous dataset for emotion analysis in text. In *Canadian Conference on Artificial Intelligence*, pages 62–67. Springer, 2011.

[8] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[9] Antoine Cully, Jeff Clune, Danesh Tarapore, and Jean-Baptiste Mouret. Robots that can adapt like animals. *Nature*, 521(7553):503–507, 2015.

[10] Taner Danisman and Adil Alpkocak. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 2, pages 53–59, 2008.

[11] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

[12] Meredith A Fox, Ru San Chen, and Clarissa S Holmes. Gender differences in memory and learning in children with insulin-dependent diabetes mellitus (iddm) over a 4-year follow-up interval. *Journal of pediatric psychology*, 28(8):569–578, 2003.

[13] Kenneth R Koedinger, Emma Brunskill, Ryan SJd Baker, Elizabeth A McLaughlin, and John Stamper. New potentials for data-driven intelligent tutoring system development and optimization. *AI Magazine*, 34(3):27–41, 2013.

[14] Gilly Leshed and Joseph'Jofish' Kaye. Understanding how bloggers feel: recognizing affect in blog posts. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1019–1024. ACM, 2006.

[15] Bing Liu. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, 2012.

[16] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[18] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[19] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[20] Kitsuchart Pasupa, Ponrudee Netisopakul, and Rathawut Lertsuksakda. Sentiment analysis of thai children stories. *Artificial Life and Robotics*, 21(3):357–364, 2016.

[21] Philip I Pavlik Jr, Hao Cen, and Kenneth R Koedinger. Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*, 2009.

[22] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[23] Francesco Ricci, Lior Rokach, and Bracha Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[24] Tom De Smedt and Walter Daelemans. Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067, 2012.

[25] Carlo Strapparava, Alessandro Valitutti, and Oliviero Stock. The affective weight of lexicon. In *Proceedings of the fifth international conference on language resources and evaluation*, pages 423–426, 2006.

[26] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4, 2009.

[27] Duyu Tang, Bing Qin, and Ting Liu. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6):292–303, 2015.

[28] Huifeng Tang, Songbo Tan, and Xueqi Cheng. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760 – 10773, 2009.

[29] Lev Semenovich Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.