Chapter 6 - Conclusions and future work

6.1
Overall conclusion

This dissertation had the objective of checking how well the overall quality (and other quality variables) experienced by the user of a SDS can be predicted on the basis of instrumentally or expert-derived interaction variables, using four different approaches (linear regression, regression trees, classification trees and neural networks). A study was made using data from experiments carried out with two SDS's with different purposes (restaurant information and smart-home control).

The results show that a high training (in-sample) accuracy does not necessarily mean high test (out-of-sample) accuracy. The PARADISE model showed that as more input explanatory variables were used, the better the accuracy would be, but this happens only for training (in-sample) data. Whenever it comes to out-of-sample (test) accuracy, this relation is different. For instance, a classification tree with five input variables has better accuracy on test (out-of-sample) data than a tree with nine input variables that include these original five.

The results from the neural networks approach are in most of the times better than the results from the other approaches. The results presented here do not prove the PARADISE model false, but shows that other experiments with other systems or with more data is necessary to make an adequate analysis. Each method for prediction has its advantages and disadvantages. A neural network can provide very good results whenever compared to the methods used until now, but the result cannot be explained, i.e. there is no coefficient like in the linear regression or node rule like in the regression and classification trees, where you can see which interaction variable is the most important and how does their variation influence the user's evaluation for instance.

A direct comparison using the same metrics between the PARADISE model and the models studied on this dissertation is presented on tables 23a, 24a, 26a, 27a, 28a, 29a, 30a and 31a. These tables show that the results from

Regression Trees and Neural Networks are better than the results from the PARADISE model (linear regression) on both in-sample and out-of-sample data.

It was recognized that some users can be well predicted and some others have terrible results. Factors like age and previous experience with SDS´s could have a correlation to this.

The neural networks model has a good potential but is pretty instable whenever it comes to initial values (different results even if same configurations and setup are used) and cannot be that well interpretable since it is a "black box".

The 'Classification Tree' method is easily interpretable and the results are good in comparison to the other methods for some of the experiments. It is a more stable procedure than a neural network, since it does not have initial values or conditions.

The results show as well, that a six months time to study this theme leaves room for more improvement, since new variables being developed (error classification) on the Deutsche Telekom Laboratories happen to have a better correlation than the majority of the interaction variables listed on this thesis. Further informative interaction variables are still needed to increase the validity of the resulting quality predictions. With the tools developed for this thesis, new experiments can be done which would give better correlation and accuracy whenever predicting overall quality experienced by the user. The following subchapter "Future prospects" has my personal suggestions for a continuation on the implementation of the methods studied in this work.

6.2
Future prospects

This section has the purpose of orientating those who plan on continuing this study. All these propositions from this section were not made due to the time restrictions of the diploma thesis.

The "Early-stop" technique used on the neural networks approach, as well as the pruning technique used on the regression and classification trees approach should be enhanced into a more effective technique.

Validation technique on NN:

Normally the validation for the early-stop technique on the NN´s should be done with independent data, so that a real prediction can be simulated. This technique is very useful, since it avoids that the network gets overfitted. The results are better than other methods tried, but these results could get even better if an independent user is used, who is somehow 'similar' to the user that is going to be predicted. The method that was tried during the making of this study is a basic way. One way to enhance it for instance, would be whenever a specific user will have its dialogue predicted, the validation data should be from a user that has the same "characteristics" of the user, like age and gender for example. The age and gender determination can already be instrumentally measured according to new techniques from speech recognition, so this wouldn't be needed to be asked in a questionnaire.

This can be seen if we use the actual test (out-of-sample) data as validation data (Neural network with three 'logsig' output neurons, 3-class evaluation):

Table 64: Statistics with different validation techniques on NN´s

Input variables	Target variable	Number of neurons on the hidden layer	Accuracy with validation on independent data	Accuracy with validation on test data
#turns, CA:#IA, UCT	Overall quality	62	41,1%	63,2%
#turns, CA:#IA, PA:CO, ts_ord, WA_iso, WPST	Overall quality	62	36,3%	67,2%
#turns, CA:#IA, WA_iso	Overall quality	62	40,1%	73,2%

On a neural network with one 'purelin' output neuron (0 to 6 evaluation) the results are better with the validation on the test data itself as well:

Table 65: Statistics with different validation techniques on NNs with linear output. The training function 'Trainbfg' was described on section 5.1.2.

Input variables	Training function	Number of neurons on the hidden layer	Pearson´s Correlation – validation on independent data	$\overline{R^2}$ – validation on independent data	Pearson´s Correlation – validation on test data	$\overline{R^2}$ – validation on test data
#turns, CA:#IA, WA_iso	Trainbfg	62	33,4%	0,07	64,3%	0,48

The validation using test data on the validation process of the early-stop technique for neural networks is like using test data as training data, so therefore these results should not be considered as test results – they were presented here to show that there is room for improvement on this area that could improve the accuracy and correlation.

On the INSPIRE System the numbers are even better:

Table 66: Statistics with different validation techniques on NN´s. The training function 'Trainbr' was described on section 5.1.2. (Input3 "'space', wa_iso', 'verb'", validation on 2 users, leave-one-out 21 times.Target value: "Use again" in 3-class evaluation – 'no', 'undecided', 'yes'.

Input variables	Training function	Number of neurons on the hidden layer	Accuracy validating on independent data	Accuracy validating on test data
'space', wa_iso', 'verb'	Trainbr	20	57,5%	**86,8%**

This means that from 61 dialogues, the neural network system could predict the right class in 53 of them.

Another approach would be stopping the training process after a fixed number of iterations or predetermined amount of CPU time, after some tests with it are performed.

New variables to be used:

The usage of new variables: the accuracy would increase for sure with variables with a higher correlation than the ones in this study (highest one on the BoRIS system→ #turns= -0,366 with 'Mean B Questions').

To prove that, a variable that has a bigger correlation to Question B0 was used (Question B1, actually a part of the questionnaire, but in theory very similar to a variable used on the PARADISE model).

Table 67: Statistics with different validation techniques and input variables on NN´s. The training function 'Trainbr' was described on section 5.1.2. (Input3: #turns, CA:#IA, WA_iso to predict Question B0 (Overall Quality). 3-class evaluation using 3 'logsig' output neurons. Test (out-of-sample) data results.)

Input variables	Training function	Number of neurons on the hidden layer	Accuracy with validation on independent data	Accuracy with validation on test data
Input3 + ts_ord	Trainbr	45	39,4%	52,9%
Input3 + **Question B1**	Trainbr	45	61,2%	74,8%

Table 68: Statistics from Experiment to predict 'Mean Questions B', using 1 'purelin' output neuron (0 to 6). Test data results.

Input variables	Number of neurons on the hidden layer	Pearson´s correlation	R^2	$\overline{R^2}$
#turns, ca:#ia, ts_ord, uct, **b1(binary)**	10	64,4%	0,412	0,395

This proves that if new variables with higher correlations are used, and improved methods for the early-stop technique on the validation process are

developed, better results can be achieved. The method itself is useful, but needs variables that are strongly correlated than the ones presented on this work. The use of emotion detection could be one of those variables for instance. So, therefore there is room for improvement on these prediction methods.

During this experiment, the new developed variable "weighted CA:IA" was also evaluated in terms of prediction capability. Its correlation to target variables was lower than the original CA:#IA variable as seen on table 6 chapter 4.2, and its results on predicting the target dependent variable were inferior than comparing with the original "CA:#IA", as the table below shows:

Table 69: Statistics on linear regression with and without the new parameter (weighted consecutive CA:IA).

Input variables	Training (in-sample) correlation	Training (in-sample) R^2	Training (in-sample) $\overline{R^2}$	Test (out-of-sample) Correlation	Test (out-of-sample) R^2	Test (out-of-sample) $\overline{R^2}$
#turns, **CA:#IA**, PA:CO, ts_ord	38,3%	0,146	0,129	32,5%	0,101	0,082
#turns, **Weighted CA:#IA**, PA:CO, ts_ord	36,1%	0,130	0,113	30,3%	0,091	0,072

But at the same time, adding the new variable as an extra input variable makes the prediction accuracy better when used on classification trees, as the table below shows:

Table 70: Accuracy comparison with and without the new parameter (weighted consecutive CA:IA) on classification trees (experiment 2)

System	Input variables	Target value	Accuracy on predicting the 3-class evaluation
BoRIS	(#turns, CA:#IA, PA:CO,WPST, IC)	Mean B Questions	54,3% (predicting "bad", "average" or "good")
BoRIS	(#turns, CA:#IA, PA:CO,WPST, IC, weighted consecutive CA:IA)	Mean B Questions	60,9% (predicting "bad", "average" or "good")

This shows that the usage of new parameters can make the prediction accuracy better in some cases.

Other approaches for the Classification trees:

The Classification trees should have more categories than 'bad', 'average' and 'good'. For instance, a classification with 'zero'(very bad', 'one', 'two', 'three', 'four', 'five' and 'six'(very good) substituting the numbers (rounding them) would be my choice and the weighted error evaluation would be more appropriate in this case than the accuracy in this case (an error between 'zero' and 'five' would have a bigger impact than an error between 'zero' and 'two').

Different approaches, other than the four presented in this thesis

Another approach that should be tried is the usage of Hidden-Markov Models. This would analyse several different other aspects of the dialogue dynamics, since it would take successive events under consideration and see their correlation with the target values.