Delft University of Technology
Electrical Engineering, Mathematics and Computer Science
Department of Interactive Intelligence

Context-based recommender system to provide cognitive support to online chat counsellors in the Helpline of 113 Suicide Prevention

Salim Salmi

Submitted in part fulfilment of the requirements for the degree of
Master of Computer Science, November 19, 2019

Thesis committee:	Prof. Dr. Willem-Paul Brinkman	Associate Professor TU Delft
	Prof. Dr. Mark A. Neerincx	Full Professor TU Delft
	Prof. Dr. Cynthia C.S. Liem	Assistant Professor TU Delft
	Dr. Saskia Mérelle	Senior Researcher 113

Abstract

STUDY OBJECTIVE: Suicide crisis chat counsellors work in an environment which demands high emotional and cognitive awareness. A shared opinion among counsellors is that as the chat conversation turns more difficult it takes longer and more effort to come up with a response. Supportive technology might resolve this "writer's block" by giving inspiration in the form of responses other counsellors gave to similar situations. 113 Suicide Prevention has an extensive chat corpus of previous conversation between help-seekers and counsellors that can be leveraged to find these situations. This thesis focuses on the design, development and experimental evaluation of a support system that aims to solve the problem of writer's block for counsellors in suicide crisis chats.

METHODS: A context-based retrieval system was build. SIF sentence embedding was used to find similar partial chats using cosine similarity. The chats in the corpus were split up and embedded using a sliding window approach, to retrieve only the relevant parts in the chat and reducing reading time for the counsellor. A within-subject experimental design (n=24) with three conditions (no support, support system, expert advice) was used to measure the system's usefulness and to test the hypothesis that the system has a noticeable difference on counsellor responses, by using experts to blindly label the responses. Furthermore, another within-subject experimental design was used to test the hypothesis that the counsellors can distinguish between partial chats provided by the retrieval system and partial chats randomly selected from the corpus. For this part of the evaluation, counsellors blindly rated suggestions on how much they related to the context of a given chat.

RESULTS: Counsellors rated the support system with a 71.04 on the SUS questionnaire, corresponding to a good adjective rating. There was variability between chats for counsellor rating for utility of the information provided by the support system. Experts noticed a significant difference between the support and expert advice conditions. Lastly, the suggestions based on the context positively affected counsellor rating of relatedness to context of the chat.

CONCLUSION: Technology using NLP techniques can provide useful information for only chat counsellors to help with writer's block. Improving the quality of the recommendations, is expected to help improve the usefulness of the system.

Preface

This work was completed in partial fulfillment of my master's degree in Computer Science at TU Delft. This project has been conducted in cooperation with 113 suicide prevention, an organization with a mission for a country in which nobody dies lonely and distraught by suicide. I would like to express my sincerest gratitude to 113 for the opportunity to cooperate with them my master's thesis. The support of the research team and the cooperation of the counsellors was indispensable. The opportunity to attend and present my work at the International Association for Suicide Prevention congress was an inspirational and encouraging experience. Specifically I would like to thank Saskia Mérelle for daily support, supervision and in general find my way in the organization.

Special thanks to Willem-Paul Brinkman for supervising my thesis project and for the guidance and feedback that I have received. Similarly all the help and feedback I received from everyone that participated in the weekly thesis group meetings from Interactive Intelligence group was incredible.

Lastly I would like to show my appreciation to my friends and family for supporting me throughout my thesis and my studies. Thank you Jeffrey, Gary and August for being great friends and for all the times you provided a much needed break. Thanks to my parents the years of support during my time as a student.

Contents

Abstract **i**

Preface **iii**

1 Introduction **1**

1.1 Motivation 1

1.2 Research Question 3

1.3 Approach 4

2 Foundation **7**

2.1 113 Suicide prevention 8

2.2 Observations 9

2.2.1 Observation methods 9

2.2.2 Task Model 10

2.2.3 Opportunities to assist 12

2.3 Envisioned Technologies 14

2.4 Focus group 16

2.4.1 Focus group methods 17

2.4.2 Chat summary generation 18

2.4.3 Inspiration 20

2.4.4 Timeline and checklist 23

2.4.5 Information look-up 25

2.5 Focused problem 27

2.5.1 Problem Definition 27

2.5.2 Requirements specification 27

3 Specification **29**

3.1 Proposed Solution: Content-based recommender support system . . . 29

3.2 Sliding window . . . 30

3.3 Document similarities . . . 31

3.3.1 Document representation . . . 32

3.3.2 Similarity Methods . . . 35

3.3.3 Comparison . . . 36

3.4 Interface design . . . 37

4 Implementation **40**

4.1 Data . . . 40

4.1.1 Word2Vec . . . 41

4.1.2 Chat data . . . 41

4.1.3 Preprocessing . . . 42

4.2 Suggestion retrieval . . . 42

4.2.1 Sliding window . . . 42

4.2.2 Similarity . . . 43

4.3 Interface . . . 44

5 Evaluation **47**

5.1 Research questions . . . 47

5.1.1 Functioning of the algorithm . . . 48

5.1.2 Noticeable difference in counsellor output . . . 48

5.1.3 Usefulness of the system . . . 49

5.2 Methods . . . 50

5.2.1 Experimental design . . . 50

5.2.2 Participants . . . 52

5.2.3 Measures . . . 53

5.2.4 Materials . . . 54

5.2.5 Procedure . . . 55

5.2.6 Analysis . . . 59

5.3 Results . . . 61
5.3.1 Performance of the algorithm . . . 61
5.3.2 Noticeable difference in counsellor output . . . 63
5.3.3 Usefulness of the system . . . 64
5.4 Discussion . . . 66
5.4.1 Functioning of the algorithm . . . 66
5.4.2 Noticeable difference in counsellor output . . . 66
5.4.3 Usefulness of the system . . . 68
5.4.4 Limitations . . . 69

6 Discussion and Conclusion 70
6.1 Conclusion . . . 70
6.2 Limitations . . . 72
6.3 Contribution . . . 73
6.4 Future Work . . . 74
6.5 Final remarks . . . 76

Bibliography 76

A OSF Form: Evaluation inspiration support system for suicide crisis counselling 82
A.1 Study Information . . . 82
A.1.1 Title . . . 82
A.1.2 Authors . . . 82
A.1.3 Description . . . 83
A.1.4 Hypotheses . . . 84
A.2 Design Plan . . . 84
A.2.1 Study type . . . 84
A.2.2 Blinding . . . 84
A.2.3 Study design . . . 84
A.2.4 Randomization . . . 85
A.3 Sampling Plan . . . 86

A.3.1 Existing Data . . . 86
A.3.2 Data collection procedures . . . 86
A.3.3 Sample size . . . 87
A.3.4 Sample size rationale . . . 87
A.3.5 Stopping rule . . . 87
A.4 Variables . . . 87
A.4.1 Manipulated variables . . . 87
A.4.2 Measured variables . . . 87
A.4.3 Indices . . . 88
A.5 Analysis Plan . . . 88
A.5.1 Statistical models . . . 88
A.5.2 Transformations . . . 89
A.5.3 Inference criteria . . . 89
A.5.4 Missing data . . . 89

B Multi-level models **90**

C Assumptions plots **92**

D Consent form **95**

List of Tables

2.1 Requirements 28

3.1 Comparison between methods for document retrieval; aside from WMD all methods used a cosine similarity to compute a similarity score 36

5.1 Model comparisons for effects of suggestion type and chat on counsellor rating (n=720) 62

5.2 Model comparisons for effect of suggestion type on expert label (n=324) 63

5.3 Confusion matrix for expert labeling of counsellor responses for the support system condition and the expert comment condition 64

5.4 One-sample T-test for counsellor usefulness ratings per support types (n=24) . . 65

5.5 One-sample T-test for counsellor usefulness ratings per chat (n=24) 66

C.1 Model comparisons for effects of suggestion type and chat on counsellor rating (n=720) 93

List of Figures

2.1 Help-seeker process flow in 113 9
2.2 Triage task model 9
2.3 Hierarchical task diagram of counsellor 11
2.4 Scenarios for information overload 18
2.5 Scenario for writer's block 20
2.6 Scenarios for conversation quality 23
2.7 Scenarios for lack of information 25

3.1 Potential interface designs 38

4.1 Support system tab in detail 45
4.2 Current chat system, provided by Livecom 46
4.3 Support system interface 46

5.1 56
5.2 57
5.3 58
5.4 59
5.5 62
5.6 63
5.7 65
5.8 67
5.9 68

B.1 Linear mixed effect model 90
B.2 Crossed random effect model 91

C.1 92
C.2 93
C.3 94

Chapter 1

Introduction

1.1 Motivation

The Dutch national institute of statistics, Centraal Bureau voor de Statistiek (CBS), reported that the lives of 1917 people came to an end through suicide in the year 2017 [1]. The same report shows that the absolute number of suicides has been increasing annually since 2007, and the amount suicides of people under 20 years old nearly doubled from 48 in 2016 to 81 in 2017. This increase in suicides among the youth is a worrisome occurrence and it has now become the leading cause of death for teenagers and young adults of age 10 to 30 [2]. Research[3] has shown that suicidal individuals, especially emerging adults aging 18 - 25, are more likely to seek help from informal online sources as their risk of suicide goes up.

To help individuals who are struggling with suicidal thoughts, helplines have been set up. There is increasing evidence that helplines are an effective method to reduce stress and suicidality in help-seekers. In an observational study, researchers found improvements in a third to half of rated crisis chat users of 113 suicide prevention [4].

Suicide and crisis helplines are a preventive measure to help individuals with suicidal behaviour with straying away from attempting suicide. These helplines are services where help-seekers can contact trained volunteers and professionals (counsellors) who can listen and assist them

with their problems relating to suicide. Historically people have been able to contact these helplines have been over the telephone, but with the advent of the internet, chat services have become increasingly popular. Compared to telephone helplines, online chat helplines have approximately the same effectiveness [4].

In the Netherlands 113 suicide prevention is a national organisation with the goal of preventing suicides. For this purpose, they created a service where help-seekers can contact counsellors over the telephone or by using their online chat platform.

In the last year over 30.000 conversations were held on 113's chat platform. There are multiple possible reasons for help-seekers to opt for counselling through an online chat rather than a tradition phone call [5][6][7][8]. First and foremost it is fully private, the help-seeker can pick and chose the location they want to communicate from without having to worry about anybody knowing what they are doing. Furthermore, compared to communication using audio, communication through text is much slower. The help-seeker has time to read the reply of their conversational partner and then to think about how they want to respond. Because there is no audio or visual feed between the help-seeker and the counsellor, the help-seeker can express any kind of involuntary emotion they want without the other person noticing, as well as the help-seeker can project the image they want onto the counsellor.

While the standardised death rate of suicides has stayed the same over the past years, the actual number of help-seekers increases yearly [1]. Furthermore the crisis helpline of 113 has had a sharp increase in people's awareness of the service, from 3% to 50%. This resulted in an increasing need for counsellors as well.

However working in the crisis line as a counsellor, talking about suicide, is an emotionally high demanding job. This is a reason for the high attrition rate in suicide crisis counselling [9]. Because the service is anonymous, the counsellor will, after the conversation has ended, most likely not get into contact with the help-seeker again. This means that the counsellor will not know what happened to the help-seeker that was in such a desperate situation that warranted a call to the crisis line. Sometimes this fact can help the counsellor move on from the

conversation. Other times it can make it more difficult, when the help-seeker is in a situation where the best thing would be to inform somebody of their location and plans, which the counsellor can not do. On top of this, at hours with high traffic some of the more experienced counsellors of 113 sometimes have to handle two chats at the same time. The triage operator, who manages incoming help-seekers, sometimes has to handle more than 3. In this case an additional person for the triage has to be on standby to assist when necessary.

There many examples of technology used to offer support in related fields [10][11][12][13], but only few that are relevant to the field of crisis counselling[10]. This thesis presents a possible approach to help counsellors perform their cognitively difficult tasks through technological support.

1.2 Research Question

The main question this research will aim to answer is as follows: How can technology provide context specific support to online chat counsellors for suicide prevention to easier execute-perform cognitive tasks? To be able to ascertain if a technology fulfills this purpose, it needs to be evaluated; before it can be evaluated, the system needs to be realised; before it can be realised, there should be a specification for it; and finally before a specification there needs to be a vision of the support system. To be able to provide argumentation for obtaining each of these, the main research question was broken down into sub-questions.

To make an informed decision on what kind of support system to explore, three things need to be define: the tasks of the counsellor, the values and goals of the counsellor, and the technology. These can be defined in three questions:

- *What are the tasks that the counsellors are trying to perform?*
- *What are the values and goals the counsellors have when performing their tasks?*
- *What are relevant technologies that counsellors can use to perform their tasks?*

- *What are the requirements for a system that can support counsellor?*

The next step, after obtaining a set of requirements for a support system, should be to realise the system. Therefore, the next set of sub questions are:

- *What possible technologies can be leveraged to realise the support system?*
- *What would the possible designs look like?*
- *What is the opinion of the counsellors on the possible designs?*

Finally the design needs to be evaluated. For this to be possible, there also needs to be some way for the counsellors to use the envisioned technology.

- *How can a prototype of the envisioned technology be build*
- *How well does the prototype of the envisioned technology support the counsellors?*

The remainder of this document will aim to answer these questions.

1.3 Approach

This section describes the approach used to answer the research questions. The steps taken roughly follow the development methodology of situated cognitive engineering (sCE) [14] which is specifically aimed at the development of supportive technologies. This methodology sets up the approach to provide a reasoning behind the design choices. This is done by, among other things, involving the counsellors in the design process. By making sure the use cases are a core part of development it ensures the eventual solution is properly situated in the domain.

To answer the first set of research questions, which aim to define what type of support system to explore, required building foundational knowledge for the design choices. First, to be able to propose possible solutions, it was necessary to have sufficient knowledge of the domain; e.g.

cognitive science, human factors, human-computer interaction design, and systems engineering; as are prevalent for cognitive engineering [15].

Second, contextual inquiry [16] was used as a means of obtaining information about the context in which the technology will function. This was done through observations with the eventual users where the user will be observed while performing their tasks and asked questions about their actions. By interviewing several users and consolidating the observations, a shared perspective can be established. These observations were performed at the 113 suicide prevent chat service, with people occupying different roles in the organization, varying amounts of help-seeker traffic and different types of help-seekers. From these observations a task model was created to map out the work-flow of the counsellor. Afterwards, several possible opportunities to assist were identified within that work-flow.

Third, possible solutions and technologies in related domains were surveyed. After the domain had been surveyed the questions regarding the needs of the counsellors were answered.

Fourth, to create a prototype a design specification was required. To answer the questions of what a design specification would look like, a scenario-based design strategy[17] was used. Scenarios are possible design solutions that are sketched in terms of high level, short stories, which focus on how the technology will operate and how the user will be interacting with it. They include an actor, the actor's actions and the outcome that is expected.

These scenarios were communicated to the counsellors using descriptions, imagery, video and story-boarding. To provide the scenarios with actors, a persona was created. Personas are hypothetical stakeholders which were created from the experience gained during the contextual inquiries. To provide the scenarios with actions and outcomes, the results of the contextual inquiry as well as the knowledge of the domain were used. Two focus group discussion were held with the direct and indirect stakeholders within 113 to gather the feedback on the scenarios and answer important questions that came up during the observations. This feedback was then used to improve the scenarios and select the scenario for which a prototype would be build. The scenario was then fleshed out and used to understand what the important values were for the counsellors.

Finally, a set of requirements were made for a prototype based on the information acquired from the previous steps. The next chapter, chapter 2, covers this approach in detail.

The second set of research questions cover how to move from the specification to a prototype. Different methods to produce the desired functionality were compared. Advantages and disadvantages were considered, and the technical approach taken was chosen based on which best fulfilled the requirements. Furthermore, several designs were proposed to counsellors. This resulted in a vision for how to realise the specification and is covered in Chapter 3. Chapter 4 covers the implementation of this specification.

Finally, Chapter 5 answers the question of how well the proposed solution solves the problem. The prototype was tested using an experimental setting where the end users i.e. the counsellors could use the system, as well as experts judging the effect it had on the counsellors. This evaluation involved several conditions with an eye on usefulness, noticeable impact on counsellor responses and performance. Finally the results of this experiment were analysed and discussed in Chapter 6.

Chapter 2

Foundation

This chapter will explore the possible solutions for a support system for a synchronous chat service for a suicide prevention crisis line. The aim of this chapter is to answer the first set of research questions:

- *What are the tasks that the counsellors are trying to perform?*
- *What are the values and goals the counsellors have when performing their tasks?*
- *What are relevant technologies that counsellors can use to perform their tasks?*
- *What are the requirements for a system that can support counsellor?*

First, some context will be given regarding the setting of 113 suicide prevention, by giving a higher level overview of the environment within 113 in which the support system was situated. Second, the methods and results of counsellor observations will be specified resulting in a task model. From these observations and task model a set of problems that have been identified. Third, relevant literature will be summarised concerning technologies for text based conversation. Fourth, these findings were validated or dismissed by feeding it back to the counsellors and other related parties in the form of scenarios of possible solutions presented in focus groups, and to discover relevant human factors involved. Finally the last section will describe which

problem was chosen to be the focus for a support system and give a set of requirements for the support system.

2.1 113 Suicide prevention

This support system was build for 113 suicide prevention. One of their activities is to provide distressed help-seekers with short term counseling through a chat service. A help-seeker can be a person contacting 113 either through a phone line or an online text messaging (chat) platform. In this thesis we will only consider the people contacting 113 through the chat platform. People can also volunteer to can become counsellors. To become a counsellor, a person needs to first follow a training program. The counsellors can work from home or at the 113 office.

113's chat platform makes use of a triage system. Depending on the traffic to the chat platform, one or two specifically trained triagist will be on standby to receive help-seekers, and be their initial point of contact. Each triagist will at maximum receive three help-seekers at the same time. Any help-seekers that arrive beyond that maximum will be put in a queue. The triage will try to asses the situation of the help-seeker. The main tasks of the triage are modelled in Figure 2.2. The objective of the triage is to make sure the help-seeker is safe for the duration of the chat. Secondly they will try to determine a particular goal for the help-seeker and counsellor to achieve during the chat they will be having. Depending on the help-seeker and the situation, setting a goal will not always be accomplished by the triagist and will have to be done by the counsellor. When the triagist has performed their tasks the help-seeker will then be forwarded to an available counsellor. Counsellors will for the majority of the time only accept a single help-seeker at a time, however there are rare cases where counsellors accept 2 help-seekers.

The triage and counselling system aims to help help-seekers with short term solutions. For long term solutions 113 also provides therapy sessions for which can be applied. These are scheduled sessions done with a psychologist, but otherwise are similarly text based chats.

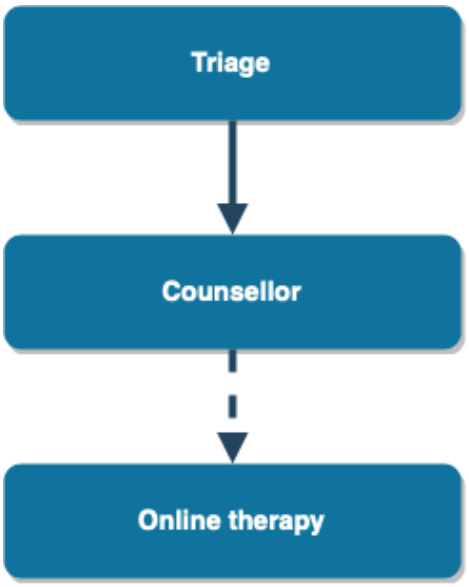


Figure 2.1: Help-seeker process flow in 113

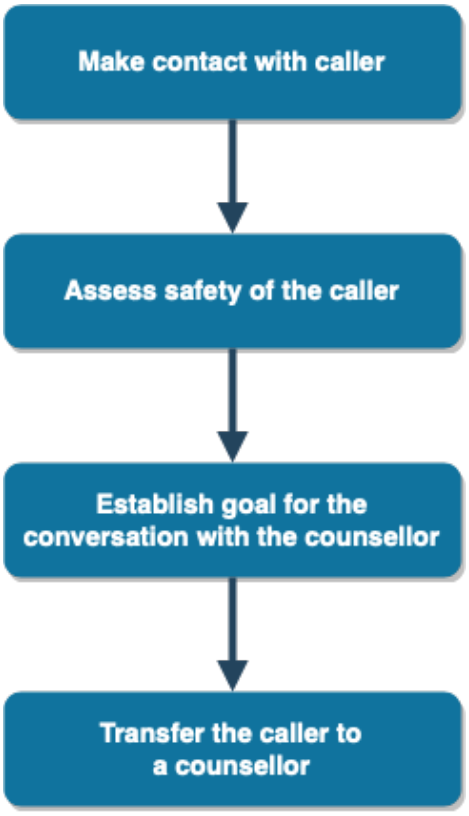


Figure 2.2: Triage task model

2.2 Observations

To be able to situate a support system in the domain of suicide crisis counseling, the domain first needs to be understood. This section will discuss in detail the approach taken to understand the domain, and will describe this resulting view of the domain. Furthermore it will identify a set of possible opportunities to assist that have been observed in the work flow.

2.2.1 Observation methods

In line with the practice of contextual inquiry [18], 8 counsellors from the suicide crisis line were observed. These observations took place before, during and after a chat with a help-seeker.

The observation was set up to be semi-structured. A couple of questions that were regularly asked were:

- How experienced are you at counselling?
- Could you describe the steps you generally take during a conversation?
- What are situations where a conversation is more difficult?
- How do you get through these difficult situations?
- What do you think of the work flow you have now?
- Do you sometimes experience stress during a conversation?

During the observation of the counsellor was observed performing their tasks and questions were posed based on observations. Several counsellors were observed as well as several triagists. During the observations digital notes were taken of the observations and answers given by the counsellor to the questions asked as well as remarks the counsellor naturally had on their tasks. Eight counsellors were observed each for a session of roughly two hours at a time and three of these counsellors for multiple sessions. Two of the counsellors were also observed performing the triage and therapy sessions. Half of the observed counsellors were experienced and the other half inexperienced.

Each session the remarks of the counsellor and their observed actions were written down. After a chat session these notes were communicated back to the observed counsellor for discussion. The notes from both the observation and the discussions were compiled to used to form a task model and identify opportunities to assist.

2.2.2 Task Model

A task model presents the work flow of the counsellors as was observed during the interviews. Figure 2.3 shows an hierarchical task diagram. After the help-seeker has been put through from the triage, the first aim of the counsellor is to build rapport with the help-seeker. To do this the counsellor needs to be aware of the help-seeker's situation. Since the triage will also

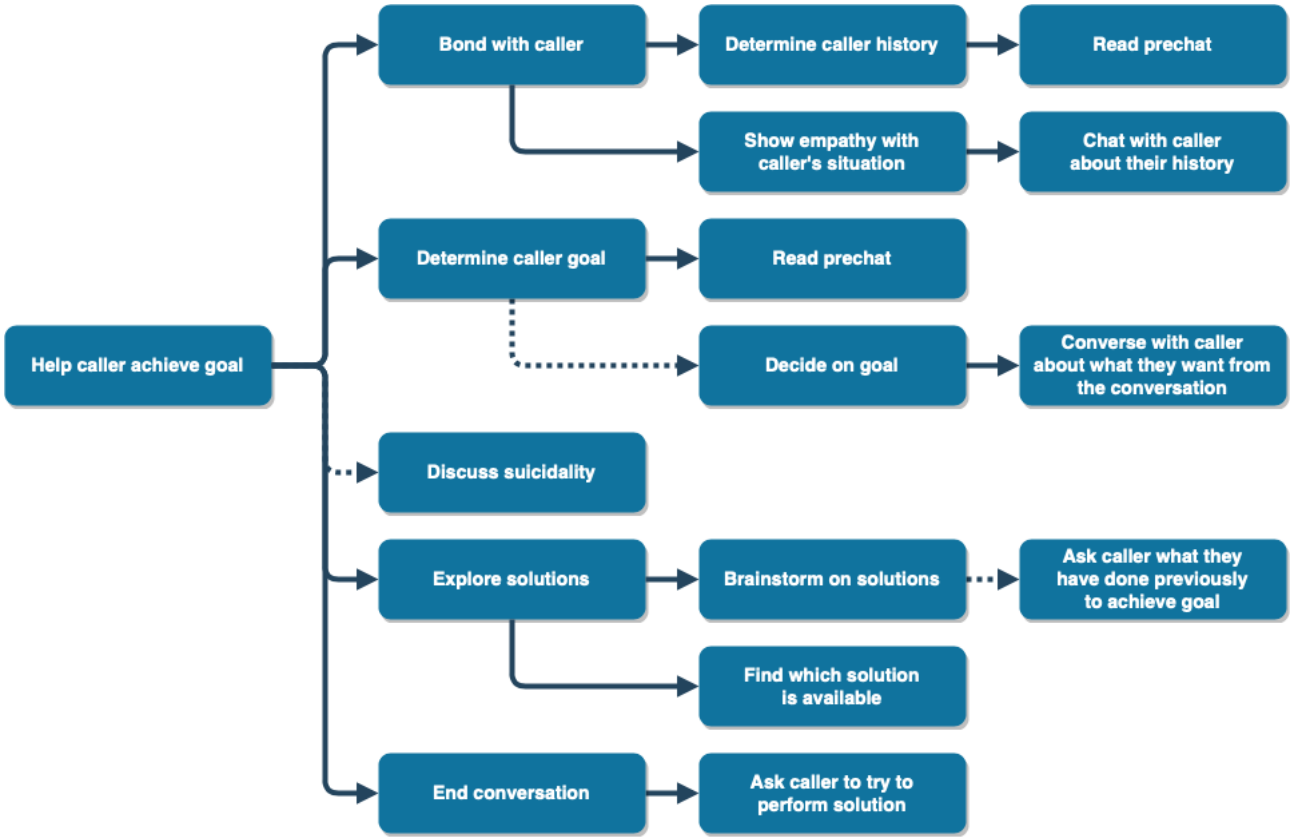


Figure 2.3: Hierarchical task diagram of counsellor

need to know the situation of the help-seeker in order to determine their safety, the help-seeker will already have mentioned it in the chat with the triage. Therefore in order to avoid having the help-seeker tell the same thing twice the first thing the counsellor does is read through the conversation between the help-seeker and the triage. When the counsellor has finished reading, they will show that they acknowledge the difficulty of the situation the help-seeker is in and follow up with additional questions about the help-seeker's situation. In a typical situation when a help-seeker requests a chat with a counsellor, the help-seeker is in that moment in distress. As mentioned in section 2.1 the regular counseling conversation only aims to help the help-seeker on the short term. Often this means that the help-seeker's goal is to calm down, however this is not always the case. The triage will, if possible, ask the help-seeker what they seek from the conversation they are about to have, in order to determine a goal for the conversation. Depending on the help-seeker and the length of the triage conversation this task is sometimes delegated to the counsellor. It can also occur that the goal that has been set is not clear enough which might cause issues later in the chat. In that case the counsellor will have to revisit the goal and make it more clear. During the conversation the counsellor will also try to find a moment to discuss the suicidality of the help-seeker. After hearing out the help-seeker the counsellor will, together with the help-seeker, look for a solution to the goal. They will attempt this in a cooperative fashion, first by exploring options. When a suitable solution has been discussed the counsellor will conclude the conversation by agreeing with the help-seeker to do whatever it is that they decided on during the conversation.

2.2.3 Opportunities to assist

Information overload

Because the 113 chat platform makes use of a triage system, a counsellor will always be added to a chat after the triage has ensured the safety of the help-seeker. This means that always some conversation has happened between the help-seeker and the triage. Often it happens that the help-seeker will already tell some information that is also crucial to the counsellor to

know. Therefore the counsellor has to read the chat between the triage and the help-seeker before the counsellor can resume the chat with the help-seeker. Furthermore, chats tended to last around 1 hour. That means that in the case the counsellor wants to refer to something mentioned much earlier in the chat or wants to address a different the next issue of the help-seeker, the counsellor will need to search for the specific mention in the chat log. This process of looking back and searching inefficiently for old information is time consuming. Trying to keep as much information stored to avoid this time consuming process unnecessarily takes up mental resources. Without having a clear overview of what happened during the conversation it is hard to maintain a proper structure. Counsellors are therefore advised to take notes during a conversation to better remember the important points they encounter during the conversation. However during the observations this rarely happened.

Writer's block

A shared opinion among counsellors is that as the conversation turns more difficult it takes longer and more effort to come up with a response. There are several aspects that can make a conversation more difficult, a few examples are:

- A non responsive help-seeker. This type of help-seeker will often ignore the messages of the counsellor. Instead the help-seeker will have a monologue until they feel that they have said their piece.
- An uncooperative help-seeker. This is a help-seeker that will listen to what the counsellor has to say but not willing to act upon the counsellor's suggestions. The help-seeker will often respond that they already tried everything or that they do not know.
- A demanding help-seeker. This type of help-seeker is looking for unrealistic results from the conversation.
- Withholding information. The help-seeker is distrustful or ashamed and does not want to talk about

- Emotional impact. The situation that the help-seeker describes themselves in is particularly dreadful.

These conversations are edge cases which require a degree of experience on the part of the counsellor to know how to deal with.

Conversation quality

A typical conversation is expected to address a certain set of points [19]. Counsellors have indicated that all these topics are not always discussed during a conversation. One of the reasons that these topics are not discussed is because the counsellor does not find an opportunity in the conversation to address the topic or simply because the counsellor may forget the topic.

Lack of information

The help-seeker demographic consist of people from all sorts of backgrounds. This means that the manner of speaking can vary from help-seeker to help-seeker. A help-seeker could not have complete mastery over the Dutch language or not speak Dutch at all. Help-seekers that use words that need to be translated by the counsellor is not a rare occurrence. The difference between older and younger help-seekers is also apparent. Younger help-seekers might be accustomed to using a lot of abbreviations and slang throughout their messages as well as keep them short and in fast bursts which can be difficult for older counsellors. Older help-seekers can write longer messages with expressions that younger counsellors might not have heard of.

2.3 Envisioned Technologies

Section 2.2.3 defined several opportunities to assist. This section describes technologies for chat based interaction, that are possible ways to approach these opportunities.

Dinakar et al.[10] created support system prototype for text-based crisis counselling called Fathom, which uses topic modelling and visualizations to provide information at a glance. During observing the work at the Crisis Text Line they identified a work-flow. They found that most of the pain points that they could identify in this work flow had to do with the taking of notes, reading of notes and reading of help-seeker information. The solution they proposed was using topic modelling with psychologists in the loop to provide visual feedback of what was being discussed in a conversation. The resulting model was then visualised using a multitude of graphs including donut and line charts. These visualizations were then included in the crisis chat interface. The results of the evaluation showed the Fathom interface as highly preferred when eliciting a list of issues and a conversation summary. However a control interface, which excluded the visualizations, was preferred during risk assessment.

The issue of forgetting the contents of a chat history over time was addressed in the paper of Tanaka et al.[20]. The proposed solution was a support system that allows the users to tag chat logs to remember past topics. The system will present candidate tags to the user based on topics from the conversation. The user will decide which tag to assign to a given conversation. Evaluation involved 12 college students, performing several sessions, either using the proposed system or a control system. The control system did not include tags from past conversations, but otherwise the interface was similar. The participants filled in a questionnaire asking about their ability to remember past conversations on a 5 point scale. The differences between the proposed and control systems was slight on most questions asked aside from the main goal: topic remembering, which had an average .6 than the control.

These works are different approaches to summarise information for chat users. Summarization is a possible method to facility keeping track of the large amount of information a chat may contain.

Related to writer's block, Sunayama et al. [21] presented a conversation support system. This support system would provide users in a conversation with topics to discuss, based on the topics they had already discussed in their conversation. Using data collected from twitter, they created clusters of related topics. The topic that was being discussed in the conversation

was then compared to the clustered topics. The most related topics were then suggested to the users to carry on the conversation. Nguyen et al.[22] proposed a recommendation system that ties into a group chat application. This chat application helps users in tourism setting to find a place of interest (POI). The chat application can provide users with suggestions for POIs, by finding highly similar POIs to the already discussed POIs. Similarity is calculated through a set of features extracted from descriptions of the POIs as well as user opinion on each POI. From an evaluation involving 15 participants, they found that users found the system usable and that the recommendations were useful. Both the work of Sunayama and Nguyen give recommendations to a person in the conversation based on the contents of the conversation. The types of recommendations, as well as how they create them differ between the two approaches, due to the differing settings of the chat applications.

Lack on information during a chat can be approached using a similar method to Watson, proposed by Budzik et al.[23], a method for just-in-time information retrieval by gathering contextual information. The proposed application extracts contextual information from a document the user is interacting with, and uses this information to query internet search engines to find relevant information for the user. In the context of a crisis chat, this application can give the counsellor quick access to information while a chat is ongoing, without the need to query a search engine manually.

2.4 Focus group

Before proposing a final solution that can be developed and tested, there are still several questions unanswered. To this end a focus group was held to get more insight on these questions:

- How do the counsellors experience the identified problems?
- What the views of the counsellors are on possible solutions?
- Would the counsellors use a solution?

2.4.1 Focus group methods

A focus group is a "controlled group discussion"[24] on a specified topic with participants specifically selected from a particular population. To facilitate the discussion and to set the topic, for each of the four problems identified in Section 2.2.3, two scenarios were created. For the purpose of generating discussion these scenarios were intentionally contrasting. The technologies described in Section sec:envisioned-technologies were used as a base to create scenarios. For this same reason each problem also has a claim made on the scenarios. Before the discussion started each participant had to fill in on a form if they agreed or disagreed with this claim. This form was not documented, but rather was used to prevent remarks of other participants from biasing their own stance. Before the discussion started the participants had time to form their own opinion on the matter and not be influenced by the responses of the other participants. Because their stance was recorded on paper they would have to stand by what their initial impression was. When the participants were done forming their stance one quick round was done where each participant would inform the group if they agreed of disagreed with the claim. Finally during the discussion the participants could defend their stance. During this part written notes were taken on the arguments and the discussion.

Two focus group sessions were performed and for each a different set of stakeholders were invited. The first focus group involved the indirect stakeholders inside of 113, e.g. trainers and supervisors. The second group involved the counsellors. Both focus groups were shown the same scenarios and claims. The first group was selected because they have the knowledge of how the helpline should ideally be operating. Next to best practices and an outside view some of the participants of the first group are also direct superiors to the second group in the organisation. To have the counsellors be able to speak freely and unbiased, they were given the opportunity to attend a session for counsellors only. The focus groups had each 10-15 people attending. Each scenario was told from the perspective of a fictional actor named Emma. Emma is a volunteer counsellor with a nursing background and is 55 years of age. Afterwards the written records were compiled and the underlying issues for each problem and what values

(a) Automated summary (b) Manual summary

Figure 2.4: Scenarios for information overload

they related to were identified. The following sections will describe the scenarios that were created for each problem, the claim that the participants were presented with, and the results of the focus group discussion.

This section will describe each of the scenarios and the results of the focus group. From the results, relevant human factors were identified.

2.4.2 Chat summary generation

The scenarios created for the problem of information overload involve summarization of the conversation. Figure 2.4 shows the visual layout for the two scenarios.

Scenario A

Emma just got back from her break after a conversation. The triage operator notifies her that a new online chat is heading his way. The chat that has been put through to Emma has a dense amount of information of a distressed help-seeker who had already started to list of their troubles at the triage. In a tab next to the chat appears a summarised version of the chat with the key points that the help-seeker mentioned. Throughout the conversation Emma can see at a glance the topics that she wants to discuss with the help-seeker. New topics that they discuss are also added to the summary as the conversation continues. As the conversation continues, Emma realises that the help-seeker is making mentions to a topic that isn't in the summary.

She has to scroll back to the beginning of the chat to read find what the help-seeker is talking about.

Scenario B

Emma just got back from her break after a conversation. The triage operator notifies her that a new online chat is heading his way. The chat that has been put through to Emma has a dense amount of information of a distressed help-seeker who had already started to list of their troubles at the triage. Emma opens a new text document and put its besides the chat window. Emma begins reading the chat between the help-seeker and the triage. When she finds something of note she will copy it into the text document. Throughout the conversation Emma can see at a glance the topics that she wants to discuss with the help-seeker. Emma adds new topics that they discuss to the summary as the conversation continues.

Claim

The automatic summary of Scenario A will be more helpful than the manual summary of Scenario B, despite it not being a perfect summary.

Results of the focus group discussion

During the discussion supervisors remarked that it should already a recommended practice for counsellors to take notes during a chat. However there is a lack of consistency with this practice. Introducing more consistency would also help the supervisors aid the counsellors more easily as they can read the summary to get the gist of the conversation. A consistency in summarizing the conversation will also benefit the structure and time that conversations take, as it avoids getting bogged down in the story and forgetting what was discussed previously.

Keeping track of that state of a conversation is important to the counsellor, however the manner in which this is done requires trust. Trust is an important issue because if the summary will

Figure 2.5: Scenario for writer's block

miss a single thing trust will be broken and the counsellor will stop relying on the summary. Furthermore at the start of a conversation, taking the time to read the chat between the help-seeker and triage is not an issue. Lastly there are certain non-verbal cues in the dynamic between the help-seeker and counsellor that can not be picked up by a summary. These cues are important to the counsellor because this is information they take into account when deciding how to appropriately respond. Ignoring this aspect could damage the connection perceived by the help-seeker. This creates a fear of missing out for the counsellor. They also have to take into account the emotional aspect of the conversation. The counsellors find it important to constantly build a stronger connection between them and the help-seeker and for which they are willing to spend extra time and effort to maintain.

2.4.3 Inspiration

For the problem of the writer's block, scenarios for inspiration were presented. Both scenarios use the same layout as can be seen in Figure 2.5.

Scenario A

Emma has been in a chat with a help-seeker for about 40 minutes now. She is feeling that the conversation has not been going well and the help-seeker is steering the conversation in

the wrong direction and Emma is taking longer and longer to respond. The help-seeker just responded with a new message. Emma does not know how to respond back and starts to ponder on possible things to say. The support system recognises that Emma is taking longer to respond. After a while a tab pops up on the right of the chat with a couple of suggestions for John to respond with. These suggestions originate from past chats. The support system tries to match the conversation Emma is currently having with chats from the past and gives responses that other counsellors gave in the most similar situations. Emma sees that the first two suggestions are inappropriate for the conversation, however the third suggestion seems like it could work. Emma copies this suggested reactions that aims to put more focus on the goal of the conversation. Before sending it he changes the content to better fit the situation.

Scenario B

Emma has been in a chat with a help-seeker for about 40 minutes now. She is feeling that the conversation has not been going well and the help-seeker is steering the conversation in the wrong direction and Emma is taking longer and longer to respond. The help-seeker just responded with a new message. Emma does not know how to respond back and starts to ponder on possible things to say. The support system recognises that Emma is taking longer to respond. After a while a tab pops up on the right of the chat with a couple of suggestions for John to respond with. These suggestions are created by experienced counsellors for a wide variety of tough situations. Emma copies one of the suggested reactions that aims to put more focus on the goal of the conversation. Before sending it he changes the content to better fit the situation.

Claim

The more specific but less consistent suggestions from Scenario A are more useful than the broader but more consistent suggestions from Scenario B.

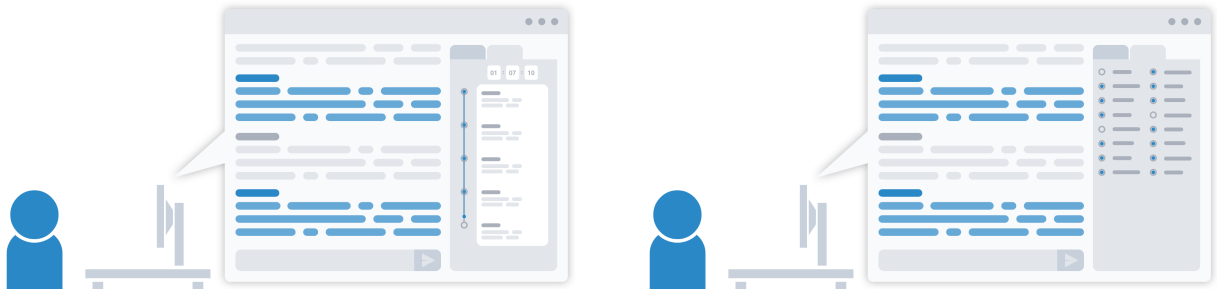
Results of the focus group discussion

The main points talked about during the discussion were that it is better to have more unique responses since as a counsellor you do not want to sound robotic to the help-seeker. Sounding robotic gives the help-seeker the impression they are not being listened to and are just taking part in a system. Much like the previous scenario this is because this can influence the connection between the help-seeker and the counsellor, which is an essential part of a conversation.

Having good rapport with the help-seeker is essential and having some element of imperfection helps in this regard. Furthermore general responses is something a counsellor can easily learn for themselves. On the other hand the focus group had concerns with the fact that looking back to responses from the past would not improve the quality for future conversation, but rather stagnate the quality. They fear that effort put towards improving the conversation quality through other means, might be affected by relying on conversations from the past.

If the counsellor would use some response given by the support system but is not aware of the reason for the response there is also the risk that the reaction does not have the desired outcome. This is due to the differences between personality of help-seekers and how they perceive responses. This influences the connection between the counsellor and help-seeker. There could also be a noticeable difference in style between the counsellors own reactions and those presented by the system. This might also affect the connection since the help-seeker might feel like they are talking to another person.

However creativity is important to the counsellors because there are a lot of different help-seekers and situations, and the counsellors will rely on one another for this. However this is not always a possibility, for example when the help-seeker is working from home, or if the floor is especially busy.



(a) Timeline (b) Checklist

Figure 2.6: Scenarios for conversation quality

2.4.4 Timeline and checklist

For the conversation quality, the two scenarios were presented using more different approaches. The main question that these scenarios aimed to ask is where the problem lied: counsellors have enough knowledge of all the elements of a good conversation but are not able to include them due to time constraints, or counsellors have trouble keeping track of all the elements. Figure 2.6 shows for each of these scenarios the visual layout.

Scenario A

Emma is nearing the end of a long shift. She gets notified of a new call from the triage. Emma knows that she is tired and looks for some assistance to make sure the quality of her conversation will not suffer. She opens the timeline tab. Here Emma sees a timeline for the major steps she needs to take in the conversation and can plan how much time she wants to spend on each topic. After 50 minutes have passed, Emma sees that there is only one step left. Thanks to the timeline she is right on schedule.

Scenario B

Emma is nearing the end of a long shift. She gets notified of a new call from the triage. Emma knows that she is tired and looks for some assistance to make sure the quality of her conversation will not suffer. She opens the checklist tab. Here Emma sees a checklist for the

major topics that she needs talk about in the conversation. She can check of each topic after she discussed them and she can gauge how many things she still needs to talk about. After 50 minutes have passed, Emma sees which topics she already discussed and which are still open.

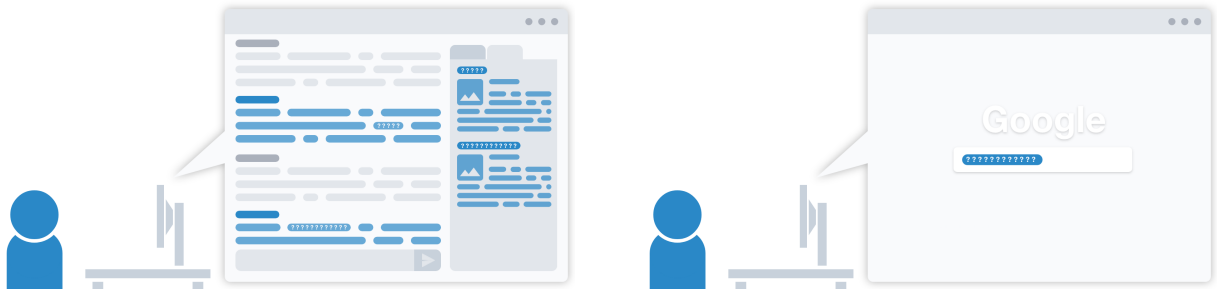
Claim

The timeline to plan a conversation from scenario A is more useful than a checklist of all the topics from scenario B.

Results of the focus group discussion

In this particular part of the focus group there seemed to be a very noticeable divide in opinion between the indirect stakeholders and the counsellors. The indirect stakeholders were more inclined towards the time management solution, as timely and well planned out conversations are in the interest of the quality provided by the helpline as a whole. Moreover, the counsellors know from experience that every conversation is different and not always every element of a good conversation will always be possible to include.

The counsellors however were mostly in favor of the checklist of Scenario B. Counsellors found that the aspect of time management was potentially stress inducing, due to falling behind schedule. Furthermore they were of the opinion that keeping track of time was within their capabilities. Having awareness of all the elements of a good conversation was favoured. To the counsellor it is more important to be able to give the help-seeker the best quality they can deliver. Having something that can help them keep track of not only reminds them of the elements but offloads some of the effort they would have to spend to keep track of the information.



(a) Automated lookup (b) Search engine

Figure 2.7: Scenarios for lack of information

2.4.5 Information look-up

The topic of lack on information on the part of the counsellor was presented with different scenarios for acquiring this information. Figure 2.7a shows an automated look-up solution, where counsellors can quickly find short answers and on the other hand an elaborate search engine capable of finding whatever the counsellor wishes to know about the language used in the conversation. The goal was to find out why the conventional method of looking up information on the internet was underused.

Scenario A

Emma is having a conversation with a young help-seeker. The young help-seeker is using slang and short hand notation for a lot of the words they are using. Emma is not very familiar with these and decides to look some of them up to better understand the help-seeker. Emma clicks on the words she does not understand and in the look-up tab a short explanation of the word shows up. Emma sees that the word she clicked on is English and can see the translation. She can quickly click on all the other words she does not understand.

Scenario B

Emma is having a conversation with a young help-seeker. The young help-seeker is using slang and short hand notation for a lot of the words they are using. Emma is not very familiar with

these and decides to look some of them up to better understand the help-seeker. Emma copies the word she does not understand. She then opens a new tab in her browser and goes to google, her favourite search engine. She pastes the word there and looks at all the search results that pop up. She sees that it is a abbreviation often used by teens. Emma can choose the source she uses to get the meaning.

Claim

The time saved by using an automated look-up is more important than being able to pick a specific source.

Results of the focus group discussion

The focus groups found that there was little reason not to use the automated look-up system, as the counsellor could always look things up on a search engine if the automated look-up did not provide a satisfactory answer. However there was again the issue that a degree trust was needed in the system for it to be used frequently.

An argument towards looking up information at all was presented. Counsellors argue that they should not always need to know everything the help-seeker throws their way. Moreover it is often desired to have the help-seeker explain things from their own point of view. This is because having the help-seeker express these things helps in building rapport since they will perceive the counsellor as being interested. Using a system that provides information and translations might play on insecurities of a counsellor as it would suggest their general knowledge and language skills are perceived to be lacking when in reality this might not be expected of them. Lastly it also influences how the counsellors are perceived by the help-seeker. For example expressing knowledge on topics they should not reasonable be informed about might make them seem less relatable. This can change the perception of the counsellor as being person to talk about their troubles with. This is turn relates to need for the counsellor to be able to build rapport between the them and the help-seeker.

2.5 Focused problem

Based on project scope and counsellor feedback, the main problem for which a support system was developed was the problem of writer's block. This section gives an in depth definition of the problem and a requirements specification for a proposed solution to the problem.

2.5.1 Problem Definition

The problems encountered by the counsellors can be categorised as one of three:

1. The situation that occurs is tough because the counsellor has no experience with dealing with that situation.
2. The situation that occurs is tough because of the emotional impact the conversation has on the counsellor.
3. The situation that occurs is tough because of a problem in the meta conversation, i.e. the conversation develops in a way the counsellor is uncomfortable with.

Each of these categories would require a different type of support. This support system will explore a solution to item 1, the problem of no experience.

Specifically the target set of problems that were examined are the problems that can be identified within a span of a few messages but for which the solution is not quickly apparent. These problems are potentially easy to solve within the space of time that a counsellor has to come up with a reaction.

2.5.2 Requirements specification

Based on the initial research question and the subsequent findings in this chapter, a list of requirements for the support system was compiled, shown in Table 2.1.

1	The support system should give the counsellor inspiration during a crisis chat to overcome a writer's block.
2	The support system should adapt to the context of the chat.
3	The inspiration the support system provides should not be generic advice.
4	The inspiration the support system provides should be conducive to a good quality conversation.
5	The inspiration the support system provides should be recognised from a few interactions between the counsellor and the help-seeker.
6	The support system should not prevent the counsellor from interacting with their current conversation, or be intrusive to their current conversation.
7	The retrieval and presentation of the information to the counsellor upon request must be at most in the magnitude of seconds.
8	The counsellor should not have to spend more than a couple minutes using the support system.

Table 2.1: Requirements

Chapter 3

Specification

This chapter builds upon the requirement found in Chapter 2 and proposes a method to this problem, to answers the research question: *What possible technologies can be leveraged to realise the support system?*. Several methods for realising the solution will be explored. This will include some relevant literature. Finally, to answer the research questions *What would the possible designs look like?* and *What is the opinion of the counsellors on the possible designs?*, possible designs of the interface are explored and discussed with counsellors.

3.1 Proposed Solution: Content-based recommender support system

The solution to the problem proposed in this chapter is to give inspiration in the form of recommendations of reactions other counsellors have given to similar situations. The problems of unfamiliar situation can often be localised to a couple of messages. This makes it suitable to judge quickly if situations match by inspecting only a few sentences. Furthermore, 113 has an extensive chat corpus of previous conversations between help-seekers and counsellors that can be leveraged to find these situations. The information of interest for the counsellor are the window within the chat where the problem occurs and what has been said after. When the

problem is quickly identifiable, the counsellor can then easily match it with other chats given the problem window. Afterwards, when a chat that matches their problem has been found by the counsellor, they can read the approach that their colleagues took to draw inspiration from and apply this in their own chat. Because time is valuable for counsellors when dealing with a distressed help-seeker, a recommender system can select a small subset for the counsellor to inspect.

Since not the whole chat is of interest, but only a fraction or *window* of a chat, the proposed algorithm scans the chats instead. The algorithm then compares the fraction that is scanned to the currently ongoing chat. This comparison results in a similarity score. The highest scoring windows will be assumed to match the problem. After finding the best matching windows, the counsellor's solution to problem found in the window is expected to follow in the next few interactions. Capturing a problem in a set of chat interactions and adding a set amount of interactions for context is, for the rest of this thesis, referred to as a "suggestion", as it tries to provide a possible way to confront the problem.

3.2 Sliding window

Because the support system will be focusing problems that are contained within a short part of the transcript and locating the exact position of a problem within a chat is difficult, a sliding window approach will be taken. The sliding window approach will map the corpus of full chat transcripts to a set of chat segments or *windows*.

$$c = \{m_1, m_2, \ldots, m_k\} \tag{3.1}$$

$$C = \{c_1, c_2, \ldots, c_l\} \tag{3.2}$$

Equation 3.1 defines a chat c of length k as a set of messages m. The chat corpus C of length l is defined in equation 3.2 as a set of chats. Using these definitions we can then define a window as:

$$w_{ij} = \{m_i, m_{i+1}, \ldots, m_{i+n} \in c_j \mid c_j \in C\} \tag{3.3}$$

$$W = \{w_{ij} \mid 0 < j \leq |C|,\ 0 < i \leq |c_j| - n\} \tag{3.4}$$

The definition for a window w_{ij} can be seen in equation 3.3, and contains a set with window size n of subsequent messages in chat c_j starting at message m_i. Equation 3.4 shows the set of chat windows W of all possible windows within chat c_j for all chats in the chat corpus C.

Given a window w' a ranking of similar windows can be made by comparing them with all windows $w \in W$ and computing a similarity score s using some similarity function $S : w, w' \rightarrow s$.

3.3 Document similarities

Section 3.2 discussed the sliding window approach to finding similar problems within a chat corpus using chat windows. This section will highlight some ways to compute similarity between documents.

The approaches that will be discussed in this section are selected based on their ability to match documents with the following assumption: The document will be a window of a chat, that is of length n where n is small number. Furthermore, the methods mentioned use an unsupervised approach, as this is the nature of the data that was available.

This section will first talk about methods to represent documents numerically as input for similarity functions. Next it will talk about some similarity functions that use these representations as input. Finally it will compare the discussed models on several metrics that have been identified to be important to the counsellors in Chapter 2.

3.3.1 Document representation

A simple method to be able to compare documents it to represent them as vectors, because there are a lot of similarity functions.

TF-IDF

Term frequency-inverse document frequency (TF-IDF)[25] is a popular[26] and simple method of representing documents. The TF-IDF embedding weights words by their term frequency in the document and divides it by the document frequency:

$$tf(t,d) = \frac{f_d(t)}{\max_{w \in d} f_d(w)}$$
$$idf(t,D) = log(\frac{|D|}{|\{d \in D \mid t \in d\}|})$$
$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,d)$$

Here $f_d(t)$ is the frequency of a term t in a document d. The term frequency promotes words that appear often within the document. The inverse document frequency discounts words that appear in a lot of different documents. This approach is an heuristic that assumes that words that appears in a lot of documents will be less relevant while words that are appear often in a few documents in a few documents are highly relevant to those documents. This assumption is based on the behavior of stop words such, which are frequently used in language and would therefore have a high term frequency, but are not very relevant.

Word Embeddings

One way to take into account the semantic similarity between different words is called word embeddings [27] [28]. This method represents words as fixed length feature vectors, where the distance between vectors inversely relates to the relevance between the words corresponding to those vectors. In the context of crisis counselling for example, the words ”medicine” and ”pills”

should have a smaller distance than "medicine" and "bullying".

There are several algorithms to compute these vector representations, like latent semantic analysis [29] and GloVe [30]. Mikolov et al. [31] developed an algorithm dubbed word2vec improving on previous methods [32]. This algorithm uses a shallow neural network of two layers that looks at co-occurrence between words. Two methods were created to produce the word embedding, namely continuous bag of words (CBoW) and skip-gram.

- CBoW: the network takes as input the surrounding words of a target word as context and tries to predict that target word. Compared to skip-gram it is faster to compute and represent frequent words better.

- Skip-gram: the network takes as input a target word and tries to predict it's surrounding words. Compared to CBoW it works better on smaller data-sets and represents infrequent words better.

Several word embeddings for the Dutch language have been made[33][34], using these methods, with differing input corpora.

The disadvantages of using word embeddings over TF-IDF are that it requires a trained set of embeddings in the language of the corpus. However unlike TF-IDF it takes into account some semantic relevance of the words.

Sentence Embeddings

Similarly to word embedding, sentence embedding aims to represent variable length pieces of text as fixed length feature vectors. A basic method that leverages word embeddings to embed sentences or documents is to take the averaged sum off all the representations of words in that sentence or document. This method works for sentences but is also applicable to documents of shorter length like the chat windows.

Resembling the method used in TF-IDF and averaged embedding, Arora et al.[35] proposed Smooth Inverse Frequency (SIF). SIF is a method that weights each word by $a/(a + p(w))$ where $p(w)$ is the estimated word frequency and a a parameter, typically set to a value between 0.001 and 0.01.

$$v_s \leftarrow \frac{1}{|s|} \sum_{w \in s} \frac{a}{a + p(w)} v_w \tag{3.5}$$

Equation 3.5 shows that the vector representation v_s of a sentence is the weighted average of the word vectors v_w of the words in the sentences. Lastly in their paper Arora et al. argue that some word anomalies cause very large components along an axis that is semantically meaningless. Therefore as a final step they remove from each embedding their projection of their first principal component. The complete embedding is show in algorithm 1.

Data: Word embeddings $\{v_w : w \in V\}$, a set of sentences S, parameter a and estimated probabilities $\{p(w) : w \in V\}$
Result: Sentence embeddings $\{v_s : s \in S\}$
forall *sentence* $s \in S$ **do**
  $v_s \leftarrow \frac{1}{|s|} \sum_{w \in s} \frac{a}{a+p(w)} v_w$
end
Form a matrix X whose columns are $\{v_s : s \in S\}$, and let u be its first singular vector.
forall *sentence* $s \in S$ **do**
  $v_s \leftarrow v_s - uu^\top v_s$
end

Algorithm 1: SIF sentence embedding

Expanding on the word2vec method, Mikolov et al.[36] presented a method of representing a variable length piece of text as fixed length vectors. It extends the CBoW model by adding a paragraph vector as input on top of the context words. This paragraph vector corresponds to a variable length piece of text and is reused for all the words belonging to that text. The resulting paragraph vector will act as a memory for the topic of the text. Compare to other methods mentioned in this section, this method learns a representation and does not use a bag of words approach. This means it does not ignore semantic information that the other bag of word methods do.

3.3.2 Similarity Methods

The methods in this subsection leverage the word representations and document representation discussed in subsection 3.3.1 to compute a single similarity score s given two representations r_d and $r_{d'}$ for documents d and d' respectively, defined as the function $S : r_d, r_{d'} \rightarrow s$.

Cosine Similarity

Cosine similarity uses the cosine of the angle between two vectors as a metric for similarity. This similarity is computed using the function shown in equation 3.6.

$$sim(a, b) = cos\theta = \frac{\vec{\boldsymbol{a}} \cdot \vec{\boldsymbol{b}}}{\|\vec{\boldsymbol{a}}\|\|\vec{\boldsymbol{b}}\|} \tag{3.6}$$

Given that a document has been represented as a vector, like with the methods used in section 3.3.1, the similarity between the documents can be expressed as the cosine similarity between the representations of those documents. This allows for easy comparison between the documents.

Word Mover's distance

The word mover's distance (WMD) [37] is a variation on the earth mover's distance. As input it takes the word representations of the words in a document. The idea is to calculate a distance or *cost* between words. The similarity between document is then seen as a transportation problem where all the words from one sentence need to transported to the other sentence, while minimizing the cost.

The cost between words i and j is expressed as the euclidean distance $c(i, j) = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2$. Using this cost, finding the distance between documents is then defined as a transportation problem with the linear program:

$$\min_{\boldsymbol{T} \geq 0} \sum_{i,j=1}^{n} \boldsymbol{T}_{ij} c(i,j) \tag{3.7}$$

Subject to:

$$\sum_{j=1}^{n} \boldsymbol{T}_{ij} = d_i \forall i \in \{1, \ldots, n\}$$
$$\sum_{i=1}^{n} \boldsymbol{T}_{ij} = d'_j \forall j \in \{1, \ldots, n\}$$

3.3.3 Comparison

Table 3.1 compares the discussed models. For several metrics, assumed performance is given, based on the literature and initial inspection by hand of the results produced by the methods. *Speed* defines how fast the method could produce a ranking, which is important for counsellors to be able to reply in a timely manner. This does not take into account preprocessing time, as this is not of concern for the counsellors working with the system. *Output* is the predicted quality of the output, based on the literature. This is important for counsellors because unrelated suggestions would take up time to read only to be dismissed, therefore the amount of unrelated suggestions should be minimised. *Hardware* is how reliant the method is on the performance of the hardware running the system. This metric was added to scope the initial implementation, and keep implementation and testing feasible.

Method	Speed	Output	Hardware	BoW
TF-IDF	Very fast	Average	Low	Yes
Avg. Word2Vec	Very fast	Good	Low	Yes
SIF	Very fast	Good	Low	Yes
WMD	Slow	Average	High	Yes
Doc2vec	Fast	Very good	High	No

Table 3.1: Comparison between methods for document retrieval; aside from WMD all methods used a cosine similarity to compute a similarity score

TF-IDF was dropped as a possible solution as the results were too limiting, compared to the other methods. WMD was dropped as a possible solution, since the computation time was

unfeasible for timely results. Average word vectors and SIF were very similar on all metrics. SIF slightly beat out average word vectors in initial testing, as well as SIF being supported in the literature for theoretically giving better output. Compared to doc2vec, the implementation for SIF was quick to compute on the hardware that was available and therefore easy to test multiple hyper parameters. Therefore, to fit in the scope of this initial exploration of the support system, SIF combined with a cosine similarity was chosen as the method of retrieving suggestions.

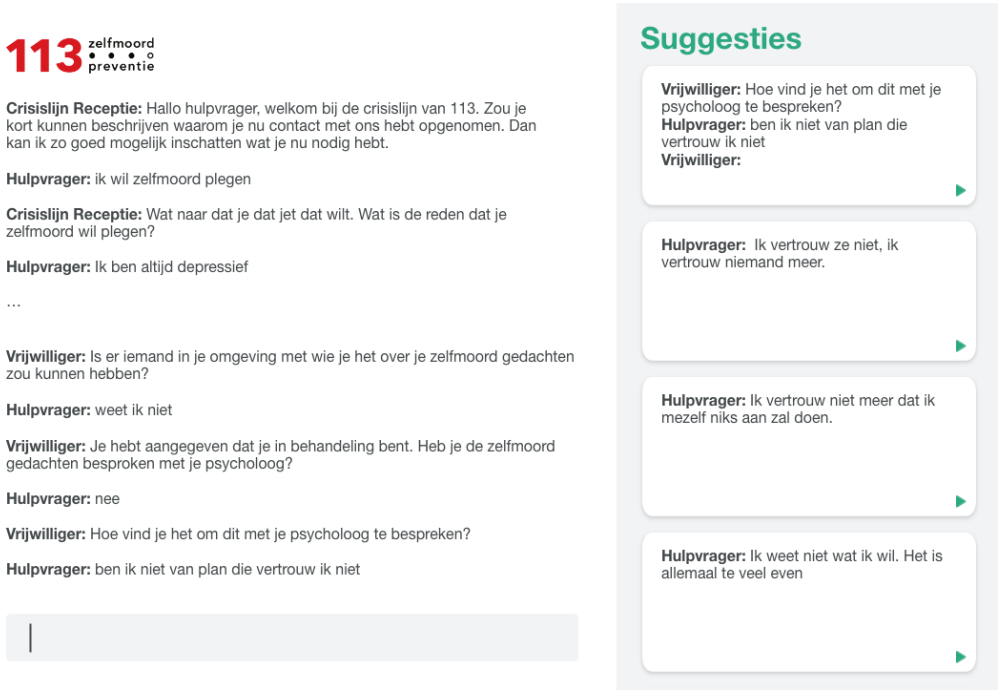
3.4 Interface design

This section will discuss how the ranked suggestions found by the recommender system will be represented to the counsellor. As is typical of recommender system, it is possible that the results wil not always perfectly match the search intent of the user. However the aim is that the system should compute a ranking where it is likely that a helpful suggestion is contained within the top n results. Because time is a luxury it is necessary for the counsellor to be able to quickly scan through a multiple suggestions to identify which ones might be relevant for them to read in more detail. The main question that had to be answered was how can counsellors quickly find the relevant suggestion they are looking for? To find an answer one on one interviews were held with 4 counsellors (2 experienced, 2 unexperienced). The counsellors were presented with mockups of contrasting designs (example shown in Figure 3.1) similar to the scenarios of section 2.4, with each design accompanying one of the following questions:

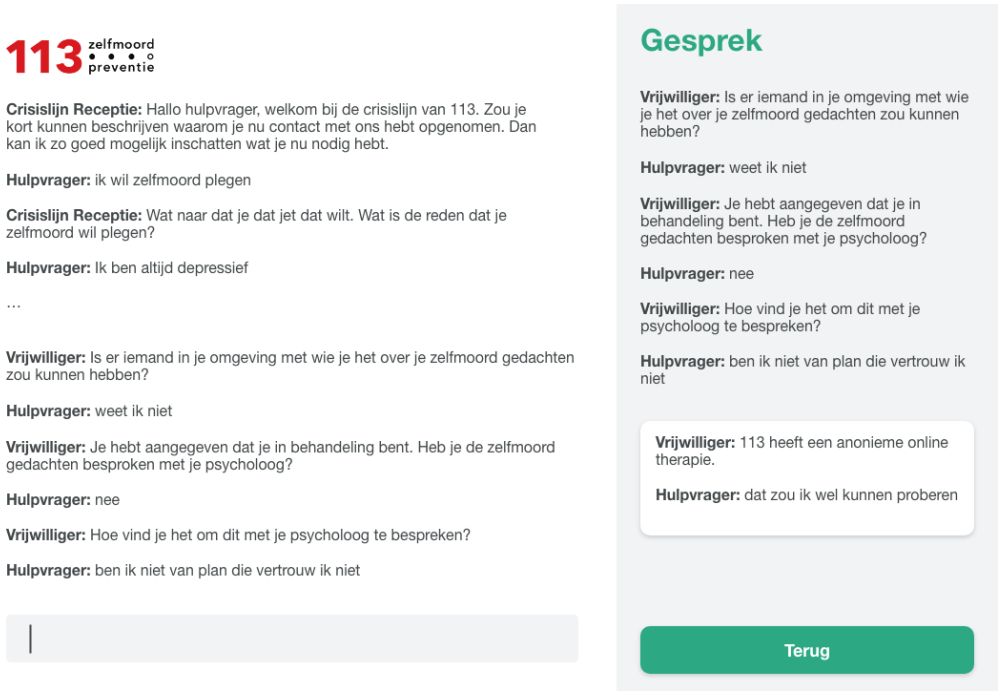
1. Would you prefer to view one suggestion at a time of a ranking of suggestions?

2. To identify what suggestion might be helpful would you rather look at the problem or at the solution?

3. To identify what suggestion might be helpful would you rather read a part of the transcript of view keywords?



(a) List of suggestions



(b) Full suggestion with context

Figure 3.1: Potential interface designs

4. Would you prefer to always have this help available and visible in the foreground or hidden and only shown when requested?

These questions were used as a base to discover what is important to the counsellor when looking for suggestions and to understand the underlying reasoning. During the interview written notes were taken. A view main points were found when analyzing the results of the interviews. Firstly there is nuance in the chat interactions can contribute a lot to the type of problem that is perceived. Secondly it is important to them to understand the help-seeker and their problem initially as to prevent a disconnect between the needs of the different help-seekers. This in turn could lead to a disconnect between the help-seeker and the counsellor, if the counsellor would give answers that don't apply to them, which makes it more difficult for the help-seeker to accept the advice of the counsellor. Lastly they feel it is important not to become to dependent on a support system because they need to be in control of what kind of conversation they are having.

Chapter 4

Implementation

This chapter tries to answer the penultimate research question from Section 1.2: *How can a prototype of the design be build?* The specification for this prototype was discussed in Chapter 3 as the content-based recommender support system. This chapter will explain how a prototype for this system was realised. The first two sections will cover the natural language processing pipeline. It starts with discussing the data that was used and how it was formatted. Then implementing the algorithm to find similar windows. Lastly this chapter will show the final design for displaying the information.

4.1 Data

This section will examine the data sets that were used to provide the algorithm with the information to search for similar windows. For this a word2vec model was used to make use of the SIF algorithm for sentence embedding, as well as the chat corpus from 113 suicide prevention. Lastly it will explain the preprocessing steps to make the chat data suitable to work with.

4.1.1 Word2Vec

The dutch word2vec model that was used was the model by coosto[38]. This model was made using Google's Word2Vec implementation using the Continuous Bag-of-Words (CBOW) model. The input was a Dutch corpus comprised of social media messages and posts from Dutch news, blog and fora. The total size of the input data was over 600 million posts and obtained between 01/01/2017 and 31/12/2017. Most other Dutch models, which were trained on properly written texts, like wikipedia and news articles. This model was trained on much more noisy data, which corresponds to the noisy nature of the chats. The embeddings in this model had a dimension of 300 and 250479 embeddings were created.

4.1.2 Chat data

To provide the the system with data, all chats held at 113, between march and september of 2018, were used.

The data was provided with the following features:

- The unaltered chat message text.
- Id of the chat message.
- Id of the chat.
- Position in the message sequence.
- Timestamp of the message.
- The source of the message with possible values:
  - **server**: Automated messages
  - **agent**: The counsellor or triagist
  - **visitor**: The help-seeker

Using the id of the chat it was possible to group all messages of the same chat. Using the position in the message sequence it was possible to properly order all the messages.

4.1.3 Preprocessing

The message text data that was provided was of the form "[<time stamp>] <user name>: <message>". The timestamp was already included in the data set as a separate column so this was removed from the message text. The username was also removed and stored into a separate field. This was done to replace the usernames in the text messages with a general counsellor and help-seeker label, to preserve privacy. Each sentence was split up into a list of words. Special characters like periods and question marks were removed, and all capital letters were replaced by lowercase letters. Lastly chats that were shorter than 20 messages were left out. This was done to limit the amount of noisy chats in the data set, that do not have enough substance to provide proper inspiration for the counsellor.

4.2 Suggestion retrieval

This section will describe the implementation of the sliding window algorithm defined in Section 3.2. First it will describe how the windows are composed and how the embedding was created. Afterwards this section will describe the how the similarity is calculated and the final selection of windows are picked and used to provide the suggestions for the counsellor.

4.2.1 Sliding window

From the text data set a sliding window data set was constructed. As defined in Section 3.2 all possible windows that fit into the chat were compiled, by using the window id and sequence id. Based on inspection of the results, the window-size that was chosen was of length 5. Server messages were left out of the windows, as well as messages with the triage.

The preprocessed lists of words for each message in the window were compiled into one single list. At this point the SIF sentence embedding would take place. The SIF embedding requires an approximation of the number of occurrences of each word in the corpus. This was obtained by tallying all the words in the data set. Afterwards per window the SIF embedding would be computed as described in Algorithm 1. The embedding for each word was looked up in the word2vec model and if not present the word would be skipped.

Each embedding for each window would result in a vector of length 300 and a total of 1286659 embeddings. The chat id and sequence number of the first message of the window would also be stored with the embedding to be able to recover the original text.

4.2.2 Similarity

After all of the steps in the algorithm that could be done before the introduction of unseen data had been done, the similarity with a new chat was computed. This is the only computational step that happens live during a chat.

Given a new chat, the system compiles the same SIF embedding for the given window. In this prototype this window is composed of the 5 most recent messages. It will then compute the cosine similarity between the 1286659 by 300 matrix of embeddings. This results in a 1286659 long vector of similarity scores.

After the cosine similarity scores are computed the list top n windows are compiled. This top n will consist of windows with the highest similarity scores, but will only include 1 window per chat. So if two windows from the same chat would have had a score that was in the top n, only the window with the best score is included in the top n. Based on initial testing and discussing with counsellors, n was set to 10 as a middle ground between having enough possible suggestions to explore, and having too much text to read through.

4.3 Interface

In Section 3.4 it was made clear what counsellors found important when narrowing down chat transcripts. These factors resulted in the design shown in Figure 4.1. The counsellor is first shown a list of suggestions. The text of the windows, which were used to match the problem in the chat the counsellor is currently having, are shown and the full text of the suggestion is hidden. The counsellor can read more of the transcript by clicking on the suggestion, if the counsellors deems that content matches the problem in the current chat. Optionally the counsellor can click the top right corner of a suggestion to hide it. The portrait aspect ratio for the container of the support system was chosen so that it could fit next to a chat, allowing the counsellor to be able to see and read their current chat and the suggestions at the same time.

To situate this design, it was compared to the interface that was in use by the counsellors. Figure 4.2 shows the Livecom[1] interface that the counsellors used. This design features the chat in the center, which takes up the majority of the screen. On the sides of the screen extra information is shown.

Figure 4.3 shows the support system in combination with a chat interface. The chat is the most important part of the interface and received a similar amount of space in this interface as with the livecom interface. The sides of the screen were used to display the extra information. The counsellor can use tabs on the top right corner to switch between other support methods like the macros, that they have available, or that could be made available in the future.

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

Figure 4.1: Support system tab in detail

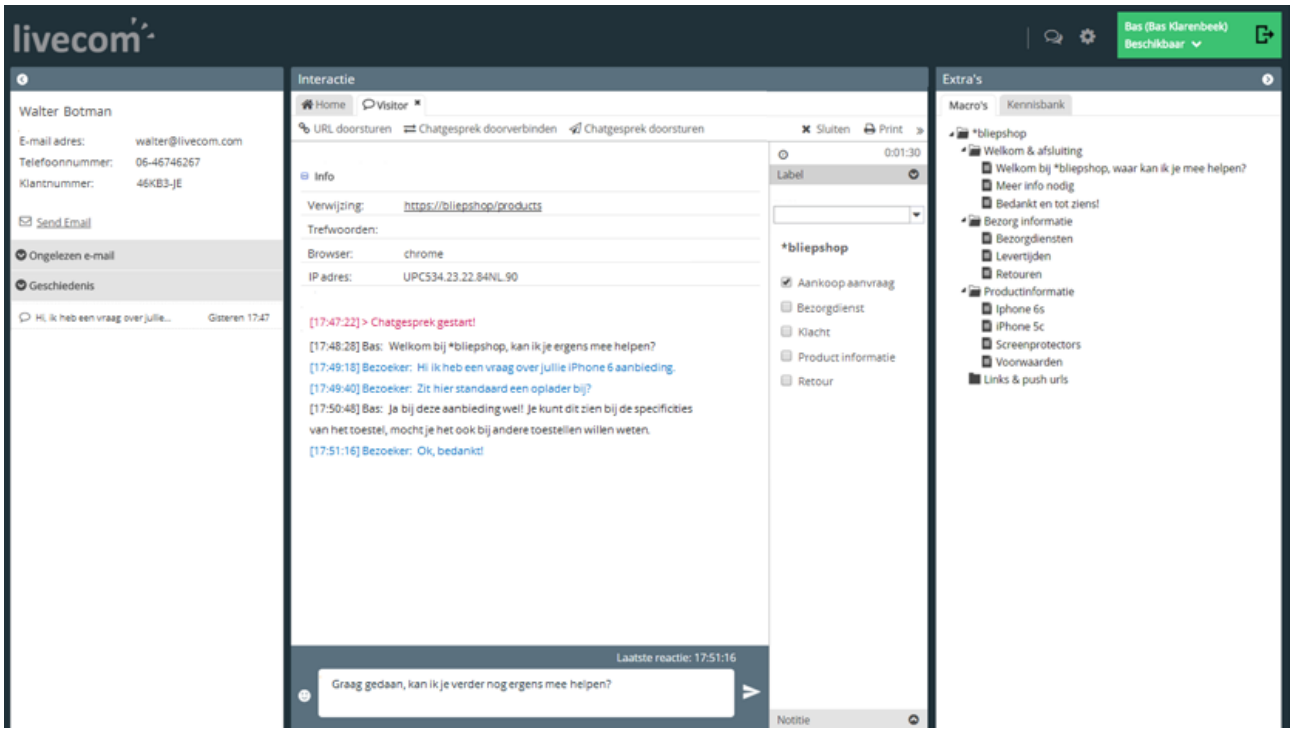


Figure 4.2: Current chat system, provided by Livecom

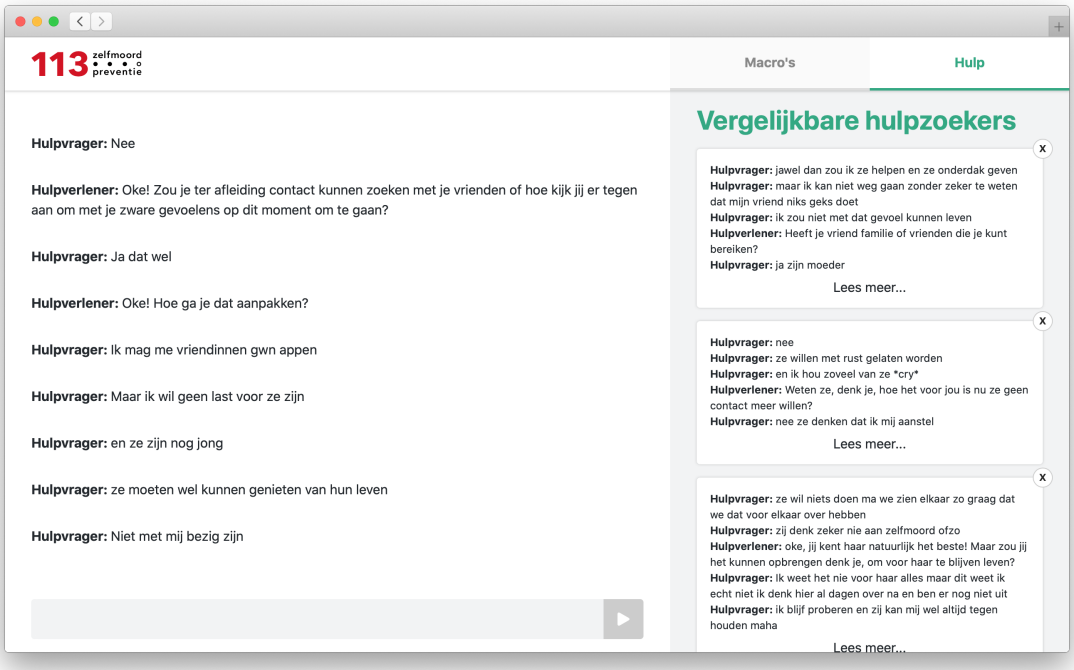


Figure 4.3: Support system interface

Chapter 5

Evaluation

This chapter answers the final sub-question from Section 1.2: *How well does the prototype of the envisioned technology support the counsellors?* This *envisioned technology* has been defined in Chapter 3, as the content-based recommender support system, and a prototype has been build in Chapter 4. To answer this question, this chapter first constructs sub questions for different components that play a role. Methods for evaluating each of these components are specified in the following section. Afterwards, the results of the evaluation are presented. Exploration of the data was done to shed light on the results. Finally the results for each component are discussed.

5.1 Research questions

Three different levels of functionality were important to the vision of the system:

- The algorithm should provide problems similar to the problem experienced in the counsellor's chat.
- The impact of the support system should be positively reflected in the counsellor's chat responses.

- The system should be useful for the counsellor.

This section specifies, for each of these levels, a concrete research question and an accompanying approach to answer this question.

5.1.1 Functioning of the algorithm

Generally, when evaluating an algorithm for information retrieval, labels on the data can be used to calculate how well the algorithm managed to retrieve the information; using statistics like accuracy, precision, recall and F1 score. In this case, the data was not labeled, however at the very least the system should be able to find matches that also contain a similar problem. Therefore, to take a first step in evaluating the algorithm, a simpler, easier to verify question was asked: Can counsellors tell the difference between suggestions related to the context of a chat, and suggestions that do not take the context into account, but have been taken at random positions from the randomly chosen chats in the corpus? The hypothesis for this research question is: *counsellors can distinguish between suggestions found by the algorithm to be related to the context of an ongoing, unfinished chat, and suggestions that have been selected from the chat corpus at random.*

5.1.2 Noticeable difference in counsellor output

For the impact of the support system on counsellor's chat responses, the ideal question to be answered would have been: do the suggestions improve the quality of the responses counsellors give. Similarly to the last component, this question could not be answered in a straight forward manner. There was no well-defined measure to judge conversation quality in suicide counselling. This, however, was needed to measure improvement of counsellor actions. As an alternative, first step towards evaluating the effectiveness of the suggestions, the focus was put on showing that there was a noticeable visible difference in the answers that counsellors gave using the

suggestions from the support system. The research question that was chosen to be explored was: Can experts differentiate between counsellor responses that have been provided either:

1. suggestions provided by the support system, based on the context of the chat

2. specific feedback from experts, who have read the chat

3. no additional help?

No additional help was added as a control condition. The expert feedback, specific to the context of the problem the counsellor is trying to tackle, was added as a comparison to the best case scenario; as a more realistic representation of the current situation, where counsellors would consult the nearby floor manager.

The hypothesis for this research question is: *Experts can distinguish between responses to a difficult unfinished chat, made with help from suggestions, help from expert feedback and no help.*

5.1.3 Usefulness of the system

The question for the usefulness of the overall system is: the users of the support system, i.e. the counsellors, should be able to use the system to reach their desired goal [39], i.e. overcoming a writer's block. The evaluation for the usefulness of the system was broken down into usability and utility [40]. The usability is how well the user can make use of the system. The utility, in the context of a recommendation system, is defined as: the value the user places on the information the system provides [41]. The question that was asked was therefore: *What is the counsellors' opinion on the usability and utility of the system?*

5.2 Methods

This section looks at the methods involved for evaluation three facets of the support system. For each of them a concrete, answerable research question has been defined; along with the approach to answer this question. Afterwards it specifies: the materials used, the participants that took part in the study, and the procedure that was followed. Finally the method of analysis is defined.

The experimental setup described in this section obtained an ethnical approval by the TU Delft Human Research Ethics Committee[1], approved 2019-03-29. Before the start of the data collection, the experimental setup was also registered on OSF[2], a framework for integrity and reproducibility of research. The full form and links to the registration, dated 2019-05-31 10:03 AM, can be found in Appendix A.

5.2.1 Experimental design

Functioning of the algorithm

To measure the algorithm, counsellors were asked to judge suggestions. This judgement involved the counsellors scoring the problem in a suggestion, against the problem in a chat that was given, based on similarity. The counsellor would give a rating on how well the problems matched. The suggestions were subject to two conditions. The counsellors were provided with three transcripts of chats, accompanied by 10 suggestions each. These transcripts stopped at a particularly difficult moment in the conversation. The two conditions of the experiment were:

- suggestions judged by the support system to be related to the context (the best matching windows, as defined in Section 3.2).
- suggestions randomly selected from the corpus, disregarding the context (control condition).

[1] https://labservant.tudelft.nl/index.php/ethics/applicant_edit/688
[2] https://osf.io/9gu2y

This experiment had within subject design with repeated measurements. This approach was taken to maximise the amount of measurements that could be taken, as reading a transcript was time intensive but for each transcript multiple suggestions could be composed. The measurement was repeated 10 times for 3 chats each, for a total of 30 measurements. Each condition appeared an equal number of times per chat and were in a randomised order.

Noticeable difference in counsellor output

The experimental setup had a within-subject design, where experts judged chat responses that counsellors gave in difficult chat situations, using different types of support. The experts judged from which support type the response originated from. The independent variable was the support type that the counsellor received. The support type had three conditions:

1. Support from support system; suggestions based on the context.
2. Support from experts; comments from experienced counsellors that have read the difficult situation.
3. No additional help.

Three difficult chat situations were used, one for each condition.

Finally, counterbalancing was used for:

- Condition order for counsellors.
- Chat order for counsellors.
- Counsellor response order for experts.

Usefulness of the system

A within-study was used with three conditions. The independent variable was the support type that the counsellor received. The support type had three conditions:

1. Support from support system; suggestions based on the context.

2. Support from experts; comments from experienced counsellors that have read the difficult situation.

3. No additional help.

Counterbalancing was used to prevent the order effect from affecting the results. Counterbalancing was used for:

- Condition order for counsellors.
- Chat order for counsellors.

5.2.2 Participants

Counsellor and expert recruitment, as well as conducting the experiment, happened at 113. 24 counsellors were each asked participate, with an average age of 27, and 21% male and 79% female participants. This amount of counsellors was chosen to allow for full counterbalancing of chat order and condition order.

Only counsellors that were interns, volunteers or trainees were eligible to participate. Counsellors that also perform the task of floor manager or therapist were not included, as they were experienced enough to come up with good answers and therefore were not in the target group of users for the support system.

The counsellors were individually asked to take time out of their counselling shifts to participate. Participation was coordinated with the active floor manager at the time, to prevent understaffing the helpline. Most of the recruitment was done in the early morning, while there were few help-seekers, and during the shift change between 3:00 and 5:00 pm, while there was a surplus of counsellors.

Each counsellor did all three components of the evaluation. Furthermore, the counsellors also did all conditions of the experiment. For the evaluation of the noticeable difference in counsellor output, this meant they created 3 responses, one for each condition.

Eight experts, average age of 35, 20% male and 80% female, were each asked judge all responses created by the counsellors, 72 responses in total for each expert. Only junior or senior psychologist were eligible as experts, as they had the most experienced background with the subject matter. The psychologist occupy management roles, as well as do the online therapy at 113. The experts did not take part in any other part of the experiment.

5.2.3 Measures

Functioning of the algorithm

The dependent variable that was measured was a score given by the counsellors on a 7 point fixed interval scale on how much the counsellors agreed with the statement "The problem in the suggestion is the same as the problem in the chat". A score of 1 meant the counsellor did not agree and 7 meant they did agree.

Noticeable difference in counsellor output

The measured dependent variable, for measuring if expert could find a noticeable difference in counsellor output, was the predicted support type judged by an expert. This variable was a three level nominal variable with the values:

- Support from support system
- Support from experts
- No additional help

To prevent expert bias, each expert would judge all of the counsellor responses.

Usefulness of the system

To measure the usability, counsellors were asked to fill in the System Usability Scale (SUS) questionnaire[42]. This is a 10 item questionnaire with a 5 point scale; from Strongly agree to Strongly disagree. SUS is widely used and the length of this questionnaire was favoured over alternative questionnaires, to minimise the participants fatiguing. The measured dependent variable for perceived utility, was a counsellor rating for the support type. The counsellors were asked the question: "How, in your opinion, did the extra information help you with finding your response?". The counsellors graded on a fixed interval scale from -3 to 3, where 3 meant the extra information was useful, and -3 meant the extra information was hindering.

5.2.4 Materials

To compose difficult chat situations, the counsellors at the helpline were asked to save transcripts of chats that contained difficult situations that they encountered. Floormanagers (senior experienced counsellors) compiled these transcripts. Six of these were eventually chosen together with a floor manager as the most difficult ones. The six transcripts were anonymised and split between two sets of three, one for evaluating the functioning of the algorithm and one for evaluating the noticeable difference in counsellor output.

Each transcript was cut off at a certain location in the chat, where it was deemed difficult to continue the conversation by the counsellors and floormanagers. The transcripts for evaluating the noticeable difference in counsellor output were read by a floor manager. The floor manager would give their comments and suggestions on how to continue the conversation. Using the last 5 messages from the transcripts, the algorithm was used to produce 10 suggestions, using the 113 chat corpus consisting of approximately 30.000 chats.

A consent form was created for the counsellors to explain the experiment, as well as ask their consent to participate and use the data they generate in this study (see Appendix D).

For the SUS questionnaire, a dutch version [43] of the questionnaire was used, to prevent

misunderstanding of the questions due to a possible language barrier.

Included in Appendix **??** are the exact questions used for the questionnaire.

5.2.5 Procedure

The experiment collected measurements from counsellors and from experts. Figure 5.1 shows a diagram of the procedure for each of them. The experiment involving the counsellors was split up into three parts. The counsellors used a test environment, in the form of a web application, that would guide the counsellor through each part. Each part was introduced with an explanation about the task the counsellor was going to performed. The measures concerning the use of the support system were taken first, as this was expected to require the most effort on the part of the counsellor. Following this, the measures for utility of the support system were taken, while the experience of using the support system was still fresh on the counsellor's mind. The measures concerning the algorithm were performed last.

1. Part 1 consisted of a simulated environment where the counsellor could read and react to simulated chat. This part was responsible for creating counsellor responses, which are needed for the measure of functioning of the algorithm. This part also recorded the measure for counsellor rating on the utility of the provided help. These two measures share the same conditions, and are therefore recorded at the same time, which reduced the amount of time counsellors had to spend reading, to minimise exhaustion.

   The simulated environment, displayed in Figure 5.2, would roughly resemble their current user interface, where the main portion of the screen was dedicated to the chat window, and with a side bar for extra information. In their current environment, this side bar they would only display macros to use. In the simulated environment for this experiment, an extra tab was added for the extra support information, called "Hulp". This extra information depended on the condition of the experiment, each can be seen in Figure 5.3. This chat would play automatically up until the point the difficult situation arose. At

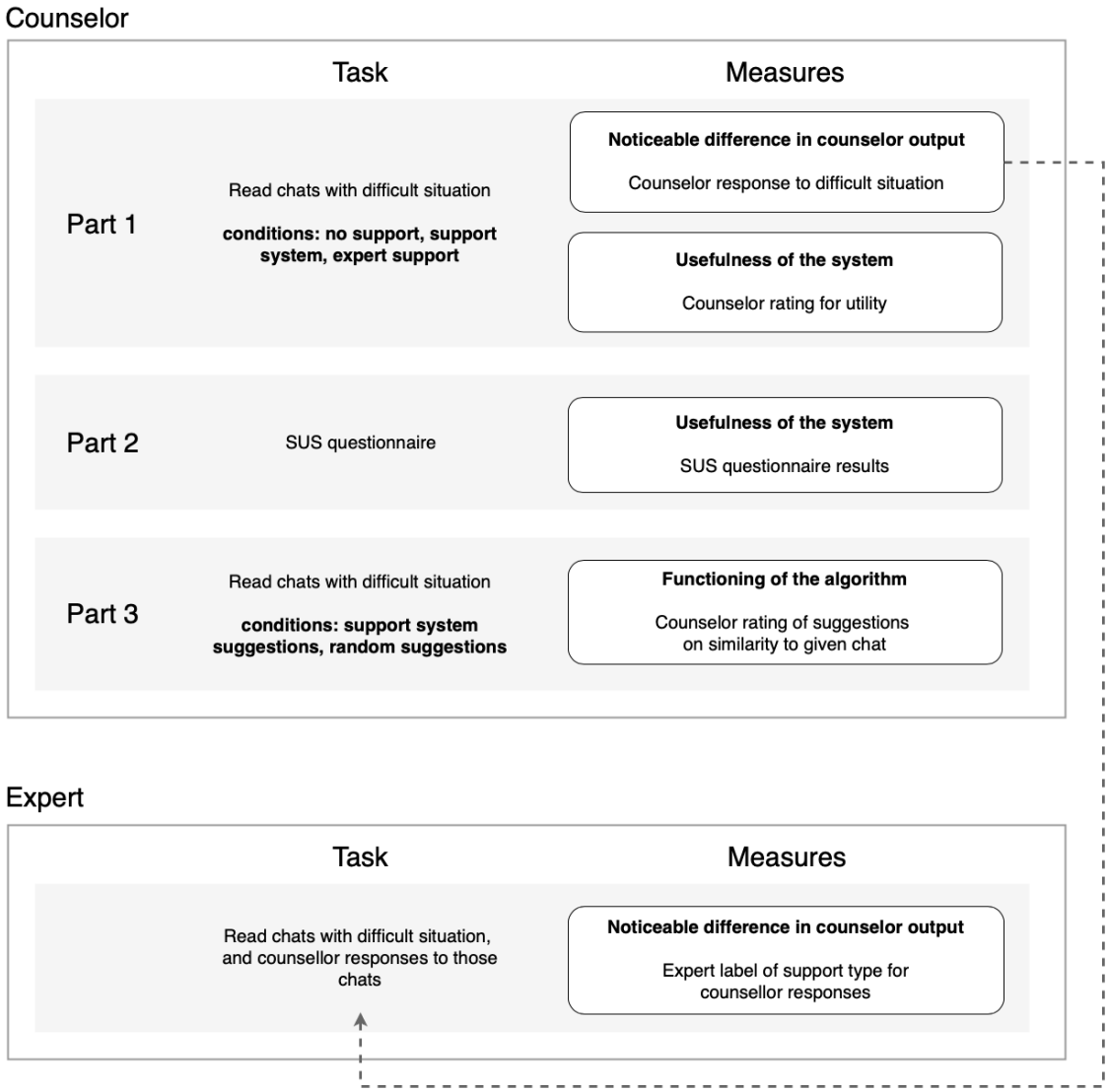


Figure 5.1: Diagram for procedure of experiment

this point a chat box became active where the counsellor could type their response to the help-seeker, in this difficult situation. This input from the counsellor was recorded. Directly after the counsellor submitted their response, they were asked to rate the utility of the support type.

2. Part 2 contained a digital form for the SUS questionnaire.

3. Part 3, seen in Figure 5.4, was used for the component of evaluating the functioning of the algorithm. The left side of the screen contained the transcript to be read. The right side of the screen contained 10 suggestion from other chats, 5 random and 5 matched by

the algorithm. Below each window was a fixed interval scale of 1 to 7, to rate how much the window related to the chat.

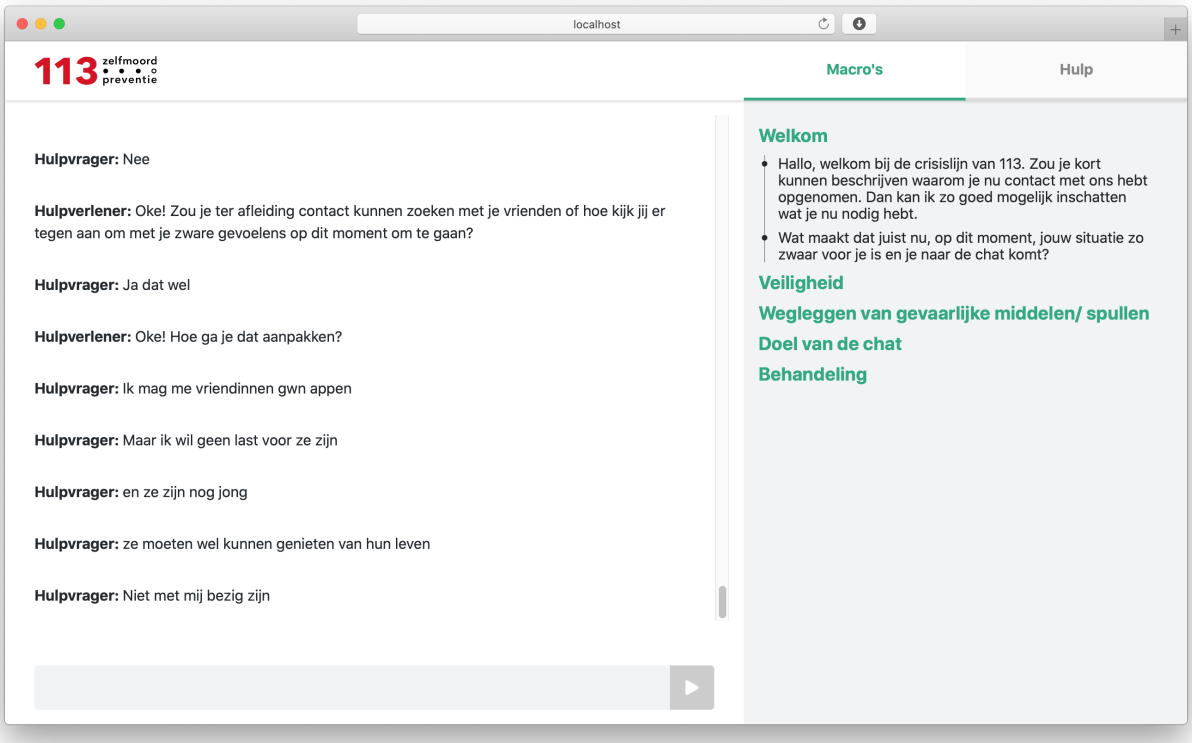


Figure 5.2: Chat interface of the experiment

Before the start of the experiment the counsellor would read and fill in the informed consent form. The counsellor then got assigned an id and an url to the experiment, which had the conditions and chats ordered based on the id.

The experiment began with a short introduction about the entire procedure. Next the counsellor read a dummy transcript, followed by a 5 minute window to explore and familiarise with the support system. No output was required from the counsellor at this time. Next part 1 of the experiment started, where three conversation transcripts that contained an unresolved difficult situation were read by the counsellor. Each new line in the chat appeared after 2 seconds until the chat reached the point where the counsellor had to reply. When that point was reached a timer of 2 minutes started and the counsellor had the opportunity to give their answer within

Figure 5.3: Conditions of the experiment: left shows expert comments; center shows support system information; right shows no extra help.

that time. This was done to give each counsellor the same amount of time to assess the difficult situation and come up with their answer. The duration of 2 minutes was chosen, in cooperation with the senior psychologists that lead the crisis line, as an appropriate amount of time that counsellors should be taking to respond to the help-seeker. During this interval the assisting information available for the condition appeared for the counsellor to use. After the counsellor filled in their response in the appropriate text box, they were asked to rate how the condition helped them in finding their answer. By completing part 1, the counsellors have had experience with the support system. They were then asked to fill in the SUS questionnaire in part 2. Finally, in part 3, the counsellors would read the 3 transcripts of the chats with difficult situations. For each transcript they were presented with 10 suggestions, and were asked to rate each suggestion on how much it related to the problem that presented itself in the chat transcript.

Data from the counsellors was collected for a period of 3 weeks, afterwards the 8 experts individually labelled the gathered responses with what condition they believed the response was made with. These responses along with their corresponding chats were printed on a form. Below each response the experts could check off one of the three options. The experts were

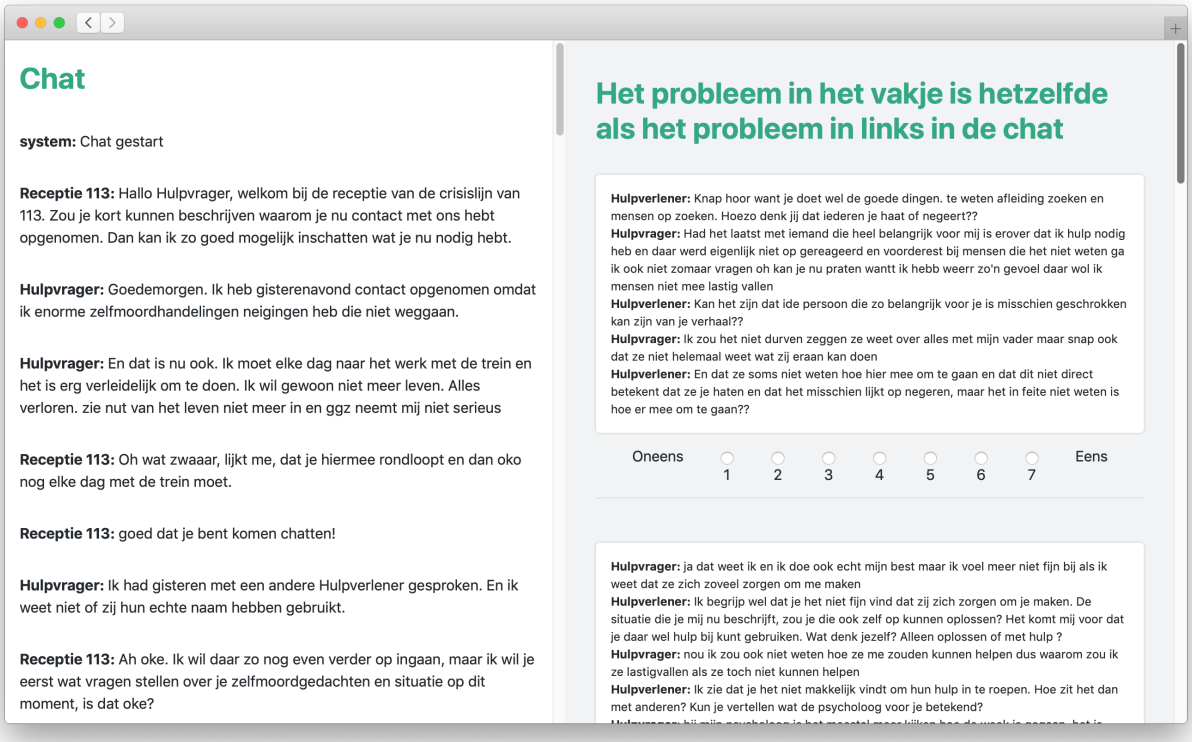


Figure 5.4: Counsellors rating each suggestion on how much they agree with the statement that the problems in the chat and the suggestion match

first given a simple explanation of the task and then committed time to judge all the response, which roughly took between 45 and 60 minutes.

5.2.6 Analysis

This subsection describes how the data was analysed. The analysis for each of the components are described separately and no correlation between the different parts was assumed. The anonymised data and scripts that were used for this analysis can be found at http://doi.org/10.4121/uuid:f65c11ff-89bf-4f65-b856-8ec773cea64f.

Functioning of the algorithm

Due to the within-subject design, linear mixed effects analysis[44] of the relationship between counsellor rating and suggestion type was performed. As fixed effects, the chat and suggestion type were entered into the model. The chat was entered as a variable because the quality of the suggestions were assumed to be dependent on the chat. As random effect intercepts for counsellors were used. This effect variable for counsellors had 24 levels. The model was computed using maximum likelihood. Results were obtained by likelihood ratio tests of the full model with the effect of the suggestion type against the model without the effect of the suggestion type.

This model assumes homoscedasticity and a normal distribution of the residuals. The resulting residuals of the model were not normally distributed. To ensure the model assumptions held, the response was log-transformed. For the log-transformed data, visual inspection of the histogram and qq-plot of the residuals did not reveal any obvious deviations from normality. Furthermore homoscedasticity was assumed from observing the scatter plot of the residuals and fitted values of the model. Because this transformation was not planned, and not included in the OSF form, the results were checked with the untransformed measures. The results for the log-transformed measures and the results without any transformation, were the same. The results reported are using the model using the measures without transformation, since this was the initial intended approach. The results for the log-transformed data and the plots that were inspected to confirm the model assumptions, can be found in Appendix C.

Noticeable difference in counsellor output

This part was analysed by comparing two conditions at a time. This resulted in three data sets, where the dependent and independent variables were binomially distributed. Therefore, a generalised linear mixed effects analysis[44] of the relationship between the dependent variable, expert label, and independent variable, support type was performed. Because multiple models were used, using a Bonferroni correction, the acceptable p value for significance was set to

0.016.

The setup for this experiment had two random effects, which do not form a hierarchy. Therefore model included a crossed random effect to account for counsellor and expert. The model was computed using maximum likelihood. A visualization for the linear mixed effects models can be found in Appendix B.

The fixed effect used was the support type (no support/ suggestion from algorithm / expert help). As random effects, intercepts for counsellors and experts were used. Results were obtained by likelihood ratio tests of the full model with the effect of the suggestion type against a null model without the effect of the suggestion type.

Usefulness of the system

For the SUS questionnaire, the overall scores were calculated. This was done by first adjusting the scores for the even numbered questions to their reverse, afterwards scaling the scores to a range of 0-10 and finally summing the 10 questions, resulting in a total score with a possible range of 0-100. These results were averaged into a mean score for the support system.

An one-samples t-test was conducted to compare the utility score to 0, as the neutral middle, for each support type. A negative deviation from 0 was assumed to indicate the information was hindering, and a positive deviation was assumed to indicate the information was helpful.

5.3 Results

5.3.1 Performance of the algorithm

Table 5.1 shows the effects of suggestion type and chat on the outcome measure of counsellor rating, for the log-transformed measures.

Suggestion type significantly predicted counsellor rating b=1.07, t=7.66, p< .0001. The suggestions from the algorithm increased the rating given by counsellors on by 1.07, from an average of 2.35 to an average of 3.42. This can also be seen in Figure 5.5, which shows the scores given by the counsellors in a bar chart. While the difference is significant, the mean of the suggestions found by the algorithm is only in the middle of the scale, with a standard deviation of 1.82. Figure 5.6 shows the distribution of the ratings for the suggestions found by the algorithm. The distribution is reasonably flat with a tendency to go down towards the higher ratings.

Lastly, the relation between between suggestion type and counsellor rating showed significant variance in intercepts across participants, SD = 0.33 (95% CI: 0.24, 0.46).

Model comparison	χ^2(degrees of freedom)	P value
Unaltered measures		
Add suggestion type	88.78(1)	<.0001

Table 5.1: Model comparisons for effects of suggestion type and chat on counsellor rating (n=720)

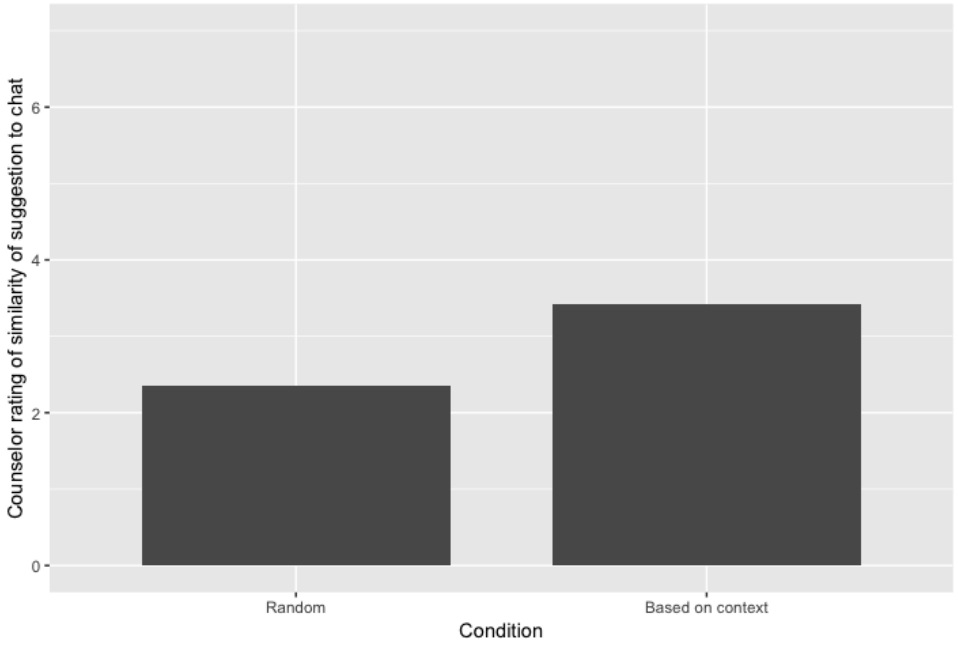


Figure 5.5: Counsellor ratings of suggestion similarity to chat.

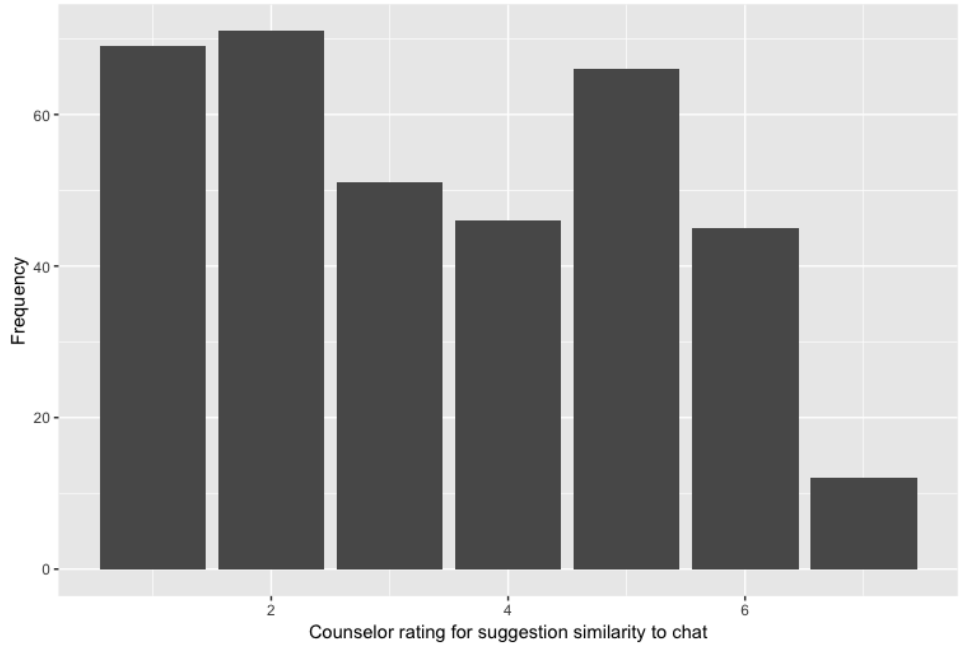


Figure 5.6: Counsellor ratings of suggestion similarity to chat, for the suggestions found by the algorithm.

5.3.2 Noticeable difference in counsellor output

Table 5.2 shows the effect of suggestion type on the outcome measure of expert label. No significant difference was found between the no support condition and any of the other conditions. Between the two conditions that received support, the support type was a significant predictor for labels given by experts (b= 0.80, z=2.62, p=0.0088). The odds the experts think a counsellor had help from expert comments is 2.23 times greater when the counsellor indeed received help from experts comments. This effect is further illustrated when looking at the confusion matrix of these conditions, found in Table 5.3.

Model comparison	χ^2(degrees of freedom)	P Values
No support vs Support system		
Add support type	1.41(1)	0.23
No support vs Expert comments		
Add support type	0.045(1)	0.83
Support system vs Expert comments		
Add support type	6.86(1)	0.0088

Table 5.2: Model comparisons for effect of suggestion type on expert label (n=324)

5.3.3 Usefulness of the system

The mean score achieved by the support system for the SUS questionnaire was 71, which corresponds to a *good* adjective rating according to Bangor et al.[45]. The scores had a 95% confidence interval of 63 to 78.

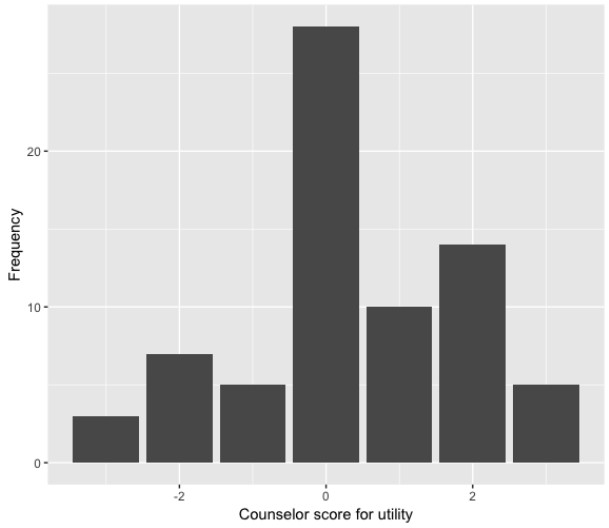
Lastly the measures of the utility ratings were analysed. The results are shown in Table 5.4: Ratings given without any support did not show a deviation from 0, the neutral position. Ratings given with the support system also did not show a deviation from 0. Ratings that do not deviate from 0 suggest that there is neither a hindering nor a helping effect experienced by the counsellors. Ratings given by counsellors that used the expert comments did show a significant positive deviation from 0, where the average rating was 1.46. This suggests that the expert comments are perceived by the counsellors as helpful. Figure 5.8 shows the utility ratings for the support types in a box plot.

Figure 5.7 shows the distributions of the counsellor ratings per condition as well as combined. While the other distributions are roughly normally distributed, the distribution for the support type condition, in Figure 5.7b, shows 2 peaks. This condition was broken down per chat in Figure 5.9, showing that chat 2 has distinctly higher ratings overall.

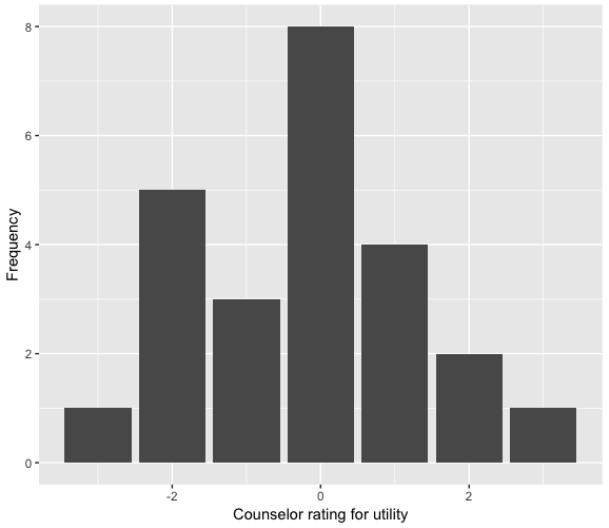
Table 5.5 shows post-hoc analysis of the utility rating per chat The second of the three chats showed a mean indicating a significant positive deviation from 0 ($p = 0.01$).

		Actual	
		Support system	Expert comments
Predicted	Support system	76	46
	Expert comments	36	47

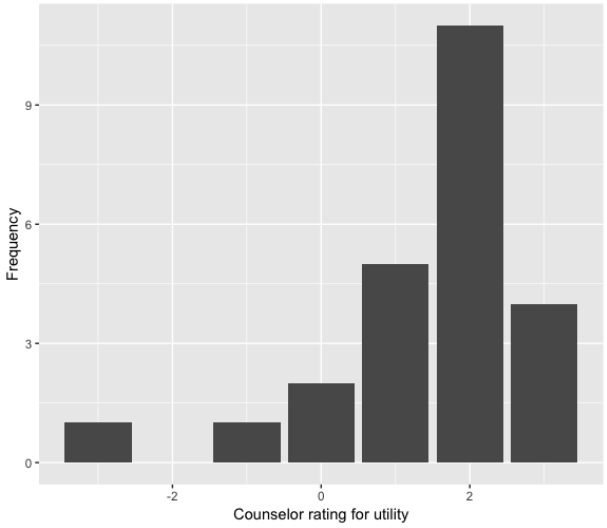
Table 5.3: Confusion matrix for expert labeling of counsellor responses for the support system condition and the expert comment condition



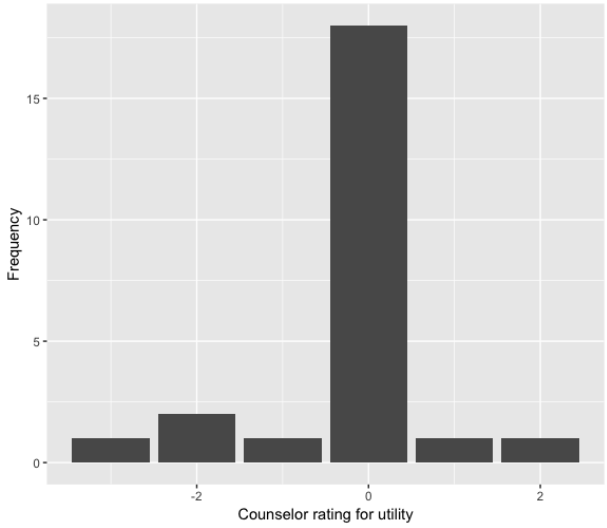
(a) All conditions



(b) Support system condition



(c) Expert support condition



(d) No support condition

Figure 5.7: Distribution of the counsellor ratings for utility

			Confidence interval				
Support type	Mean	SD	Lower	Upper	t	DF	P Values
No support	-0.21	0.95	-0.62	0.20	-1.04	23	0.31
Expert comments	1.46	1.91	0.87	2.04	5.17	23	<0.001
Support system	-0.21	2.26	-0.84	0.43	-0.68	23	0.50

Table 5.4: One-sample T-test for counsellor usefulness ratings per support types (n=24)

			Confidence interval				
Chats	Mean	SD	Lower	Upper	t	DF	P Values
Chat 1	0.17	1.52	-0.48	0.81	0.54	23	0.60
Chat 2	0.79	1.38	0.21	1.38	2.81	23	0.01
Chat 3	0.08	1.59	-0.59	0.75	0.26	23	0.80

Table 5.5: One-sample T-test for counsellor usefulness ratings per chat (n=24)

5.4 Discussion

5.4.1 Functioning of the algorithm

The hypothesis for this component was that the counsellors can tell the difference between suggestions related to the context of a chat, and suggestions that do not take the context into account. The results for the functioning of the algorithm were as expected. The counsellors could indeed tell the difference between the two conditions. The findings suggest that text compared, using sentence embedding, are seen as related by counsellors. However, the ratings given by the counsellors, on if they agree that the problem in the suggestion is the same as the problem in the chat, suggested that while the counsellors thought the problems were similar, they rarely agreed the problems were the same, as shown in Figure 5.6.

5.4.2 Noticeable difference in counsellor output

The hypothesis for this component was that counsellors can tell the difference between no additional help and the other two conditions. The results for the noticeable difference in counsellor output experiment were inconclusive. The expectation was that the no support condition would be distinguishable from the expert support case, as these were the assumed as the worst and best case scenarios respectively. If there is no difference between these two conditions then it was expected that there would not be a difference between any two conditions. However there was a observable difference between the expert help condition and the support system condition.

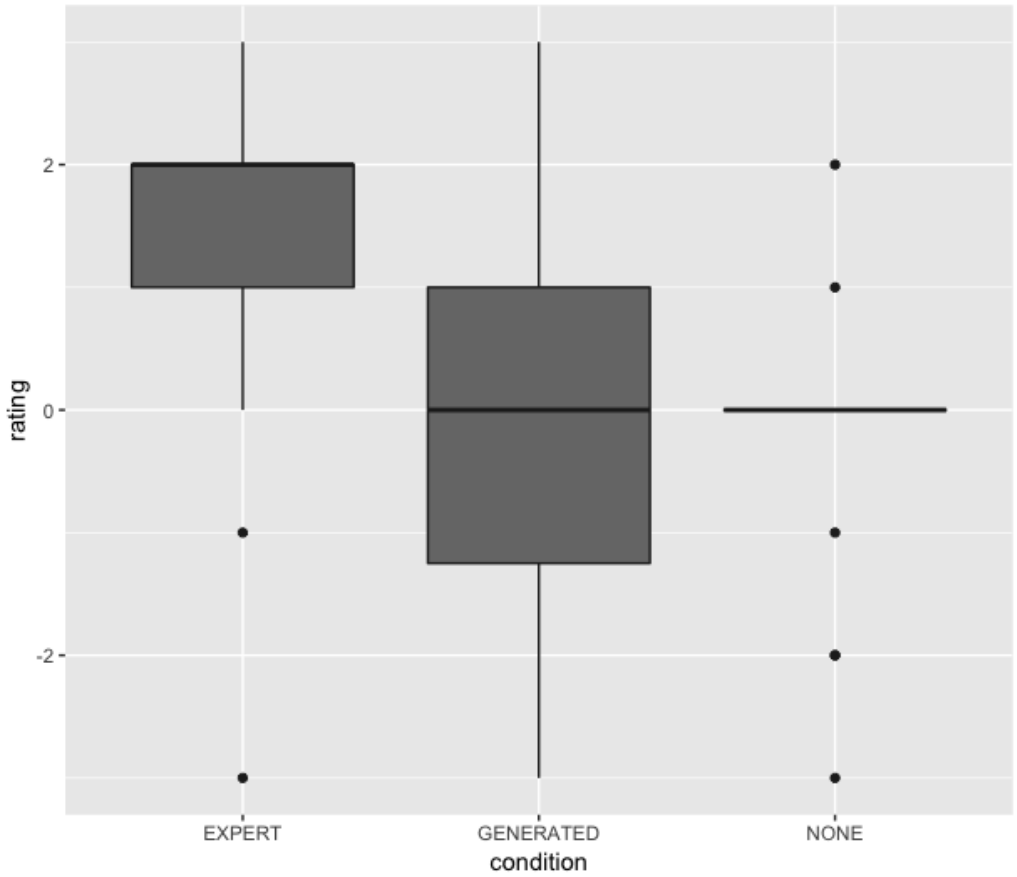


Figure 5.8: Box plots of the utility ratings given by counsellors per support type

One possible explanation for the result that there was no recognizable difference between no support and any of the other conditions is that the chat problems might not have been difficult enough for the participants. Another possible explanation could that counsellor responses without support were very varied, which could be a confusing factor for the expert judges. The counsellors that got support were all given the same information, for which it could have been possible to find a common thread in the responses. The counsellors without support, however, had to rely on their own experience and imagination.

Regardless, when given support, there is an observable difference in the type of responses counsellor give. This shows that the information counsellors get, influences the way they will respond.

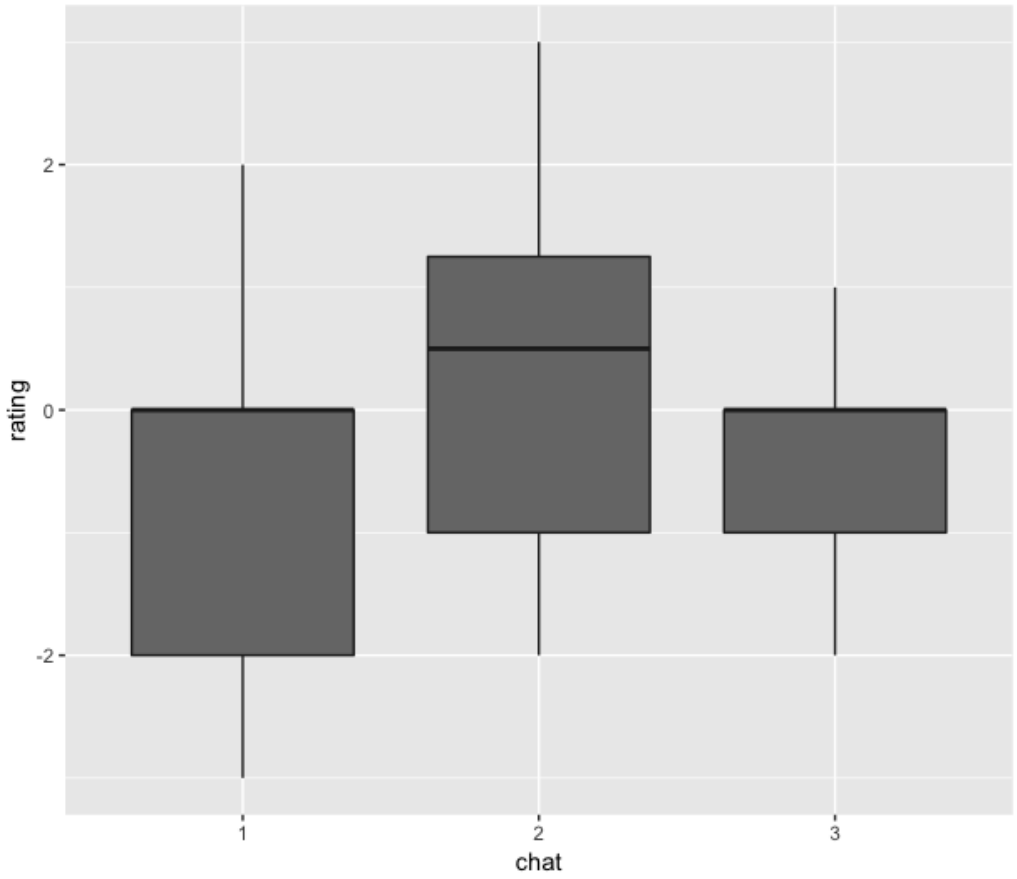


Figure 5.9: Box plots of the utility ratings given by counsellors per chat for support system condition

5.4.3 Usefulness of the system

As expected, for the utility, the highest average scores were obtained by support from the expert, with some amount of variation. This was targeted human expert advice and the best case scenario for support. Also as expected no extra support was rated 0 by nearly all counsellors, with very little variation.

The results for the support system were interesting since the opinions were divided as seen in the high variation in the ratings. One possible explanation is that, when looking at individual chats, the support system gave more useful information for particular chats. Namely the mean of chat 2 was almost a full point higher than the other two. Upon inspection, the suggestions seemed better matches for chat 2, compared to the suggestions for chats 1 and 3. Furthermore, chat 2 might have been seen as more difficult as the ratings for this chat type were higher than

the other two chats.

Nevertheless, contrasting to the no support condition, multiple cases of counsellors rating the support system useful or hindering, caused a high variability. This shows that there is room to grow for the support system, by reducing the cases where the support system is hindering, and promoting the cases where the support system provides useful information. Possible approaches for this goal are: improving the usability of system, indicated from the SUS questionnaire; Improving the quality of the algorithm, as discussed in Section 5.4.1.

5.4.4 Limitations

For this experiment, a writer's block was forced through giving a counsellor a conversation with a situation that previous counsellors found to be difficult. However, not every counsellor will have problems with the same types of situations. A solution for this limitation would be to use the support system in a real online crisis chat and wait for a writer's block to naturally occur, and record measures afterwards. It was decided not to evaluate the support system in live situations due to several constraints. The ethical issue of using an untested system in a crisis situation was an important constraint, since there was a possibility for the system to be hindering during a real crisis chat. Furthermore, implementation time on the part of the platform proprietor would have pushed the evaluation out of the scope of this project. Lastly, due the nature of live testing, every conversation would have been different. Therefore it would not be possible to have the expert comment condition, as it would be unfeasible to measure and record.

It was not possible to directly give a measure of quality to counsellor responses using the support system, as this measure does not exist. Also comparing responses with each other, to tell which one was better, was something an expert was not able to do.

Chapter 6

Discussion and Conclusion

This chapter concludes this thesis, by recapping the answers to the research questions. Furthermore, this chapter covers the limitations that were found during the course of developing and evaluating the prototype are highlighted. Afterwards, the contributions of this work are discussed. Finally, potential approaches for improvement will be recommended for future work.

6.1 Conclusion

The primary research question for this thesis was: *How can technology provide context specific support to online chat counsellors for suicide prevention to easier execute-perform cognitive tasks?* To answer this research question, it was broken down into several sub-question.

What are the tasks that the counsellors are trying to perform?
The counsellors' primary goal is to help people in a suicide related crisis situation and act as first aid to help people to deal with their crisis at that moment. They do this by determining a goal with the person in crisis, building rapport with the person in crisis, exploring solutions together and finally helping the person in crisis come to a decision on how they will

What are the values and goals the counsellors have when performing their tasks?
The way the counsellor fulfil their tasks is by building rapport with the person in crisis, so

that they have an easier time communicating with the person in crisis. Trust is therefore an important value for the counsellors. They created trust through listening to the person in crisis, and being understanding of their situation. The counsellors aimed to appear engaged with the person in crisis to show that they were listening. Lastly, authenticity also helped the counsellors appear understanding, as they did not want to appear robotic to the person in crisis.

What are the requirements for a system that can support counsellor?
The focus of this thesis was put on one specific challenge, namely a counsellor encountering a writer's block during a conversation. For this challenge requirements were compiled: The system should be able to help the counsellor through a writer's block. The system must be efficient to use and fast to interact with. It should not take up more than a few minutes of their time to find the information they are looking for, as this would conflict with their goal to be engaging. The information to help the counsellor overcome their writer's block must not be generic advice that could apply to a lot of situations, as this would conflict with the goal to be authentic.

What possible technologies can be leveraged to realise the support system?
In contextual information retrieval systems[21][22][23], contextual information, originating from the application users were interacting with, was used to build queries without the need for additional user input. Users were provided with information without the need to manually query. This concept was used, together with natural language processing, to provide counsellors with a system that can help them with a writer's block. The support system provided counsellors with partial chat transcripts that dealt with a similar situation as they were experiencing. To find these partial chat transcripts, sentence embedding, using smooth inverse frequency[35], was used to scan chats for semantically similar text.

What is the opinion of the counsellors on the possible designs?
The counsellors want to have a design where they are able to quickly identify if a recommended conversation, and the counsellors' conversation, are experiencing comparable problems. Furthermore, counsellors were interesting in the entire context of the interaction rather than a

summary. They found that the nuance in the chat interactions was important to properly understand the situation.

How can a prototype of the envisioned technology be build?
A prototype was build using a dutch Word2Vec model along with smooth inverse frequency to create the sentence embedding. A sliding window was used to scan the chats and compare their embedding, through a cosine similarity, to the chat the counsellor was having.

How well does the prototype of the envisioned technology support the counsellors?
Evaluation of the prototype showed that counsellors found the partial chat transcripts, given by the support system, matched the problem they were having better than randomly selected partial chat transcripts. Furthermore, evaluation showed that counsellors' perceived utility of the support system varied between chats. For some chats the perceived utility was positive, while for others it was slightly negative. Furthermore, when given support, experts could tell which type of support was given, between support from the support system and support from an expert. However experts could not tell the difference between responses that received no support and responses that received support, from either of the support methods used in the evaluation. These results implied that the support system had an influence on how counsellor responded to the conversation.

6.2 Limitations

Some limitations have been identified for this research. First, the specification, development and evaluation were done in the context of counsellors for 113 suicide prevention in the Netherlands. There might be differing factors influencing counsellor behavior, and their need for support, between different organizations. Second, the support system only targets the writer's block, and specifically a writer's block caused by a problem that can be traced to a few messages that the help-seeker said. The implementation for the support system was not modelled for a writer's block created by problems in the meta conversation, or a writer's block caused by the emotional

state of the counsellor. Third, because the data was not labeled, the recommendation algorithm was not evaluated through deterministic methods such as calculating the accuracy, precision, recall and F1 score. Being unable to compare models in terms of these statistics makes it harder to determine what the best model is. This is true, not only what type of representations to use but also what hyper parameters to use. Similar to limitations of contextual document querying, Budzik et al.[23] pointed out that the results the support system provides might be similar but that does not mean they will always be relevant. Fourth, the support system only provides support during online chats, and not during phone calls. This is due to the different nature of phone calls. Not only would calls require speech to text translation, but, compared to text chat, interactions over the phone are much more immediate. Therefore there is little opportunity to use the current implementation of the support system for calls, as the counsellor would have no way of reading chat transcripts and responding in a timely manner. Lastly, the support system worked the most optimal when the system could produce the useful examples, to approach a problem that caused a writer's block for the counsellor. However it could not do this for every difficult chat situation. This is a consequence of several factors: Aside from removing chats that were of shorter length, the chat corpus was largely kept unfiltered. This meant that there were noisy chats included in the data set. The noisy chats increase the chance that counsellors encounter partial chat transcripts that are not useful, which adds to the time the counsellor needs to spend with the support system. Since counsellors' time is valuable, every partial chat transcript they read that is not relevant creates a negative experience.

6.3 Contribution

Compared to other support system for chat such as [21], [22] and [46], this corpus-based approach, for combating writer's block, through inspiration from other chats, is a novel approach in the field chat support systems and of suicide counselling. The main contribution of this system is that it can provide counsellors with a potential alternative to getting inspiration from a floor manager or a peer in the case this is not an option for them, such as during peak

hours or shifts at home. Getting partial chat transcripts from a support system could boost the confidence of counsellors, knowing they have a fallback at all times. The counsellor can discover new approaches to problems, that they themselves might not have thought of, or they can confirm their own approach by checking if other counsellors have taken the same approach.

6.4 Future Work

In terms of usability, different designs could be explored and compared against each other as a way to discover ways to more easily get the information that the partial chat transcripts provide to the counsellor. For instance, the suggestion could be summarised into a few keywords, for the counsellor to skim over. The full transcript could then be accessible by selecting a suggestion that has interesting key words. Furthermore, some examples of additions that could be explored to help the current design are:

- Visually aids, by means of accentuating important words or sentences.
- Filters that the counsellor can use to narrow down the results.
- Ability to input extra words to help the recommendation algorithm discover suggestions that rely on a broader context.

One of the points that was concluded from the evaluation is that there is an indication that the quality of the information provided by the support system influences the counsellor's perceived usefulness. There are several ways this first version the algorithm could be improved upon in a future version of the support system:

- The support system scanned chats, for comparable situations, by taking a fixed number chat messages at a time, and combining them into a "window" to compare to the current chat. Learning the representations of the windows through a machine learning method like Doc2Vec[36], could improve the quality of the representation. This is a state of the

art method that can potential better leverage the large corpus of chats that is available. Unlike the method used for the current implementation of the support system, the doc2vec method does not use a bag of words model. This adds extra semantic information to the representations, by being sensitive to the order of words. Increasing the quality of the window representations is expected to improve the quality of the partial chat transcripts.

- Clustering methods could be used to create more semantically meaningful and dynamically sized windows. This can help by more accurately capturing the problem. In this this could improve the support system's ability to find better partial chat transcripts, but would also help the counsellor in faster recognizing a partial chat transcript's problem matches their own problem.

- A curated corpus of chats that are found, by counsellors and floormanagers, to handle the problems encountered in the chat well. The result would be a curated pool of chats that cover a wide variety of problems, while also containing solutions that are regarded as proper ways to handle the problem. This can greatly reduce noise in the chats to prevent counsellors from getting partial chat transcripts that do not lead to answers, as well as the potential quality of the answers.

- Labelling chats with problem labels if they contain a certain type of recurring problem would greatly improve the quality of the partial chat transcripts. While it is likely unfeasible to label the entire corpus with all possible problems, semi-supervised learning can be used leverage potential additional information that could improve on the current unsupervised method. Furthermore, using active learning, counsellors can help in the labelling process by labelling specifically selected windows that are expected to have a higher chance of containing a problem.

6.5 Final remarks

This thesis showed a possible way to support counsellors in suicide crisis lines by assisting them in overcoming a writer's block. It showed that this type of support has the potential to have a positive influence on the counsellor's experience. Suicide counselling can be a mentally taxing job, and this technology can provide them with additional tools to overcome some of the hurdles in their journey to help people in crisis.

Bibliography

[1] Centraal Bureau voor de Statistiek, "Yearly statistics of suicides in the Netherlands." https://www.cbs.nl/nl-nl/maatwerk/2018/27/maatwerk-zelfdodingen. [Online; accessed November 19, 2019].

[2] Centraal Bureau voor de Statistiek, "Cause of death statistics in the Netherlands." https://opendata.cbs.nl/dataportaal/#/CBS/nl/dataset/7052_95/table. [Online; accessed November 19, 2019].

[3] A.-L. Seward and K. M. Harris, "Offline versus online suicide-related help seeking: changing domains, changing paradigms," *Journal of clinical psychology*, vol. 72, no. 6, pp. 606–620, 2016.

[4] J. K. Mokkenstorm, M. Eikelenboom, A. Huisman, J. Wiebenga, R. Gilissen, A. J. Kerkhof, and J. H. Smit, "Evaluation of the 113online suicide prevention crisis chat service: outcomes, helper behaviors and comparison to telephone hotlines," *Suicide and Life-Threatening Behavior*, vol. 47, no. 3, pp. 282–296, 2017.

[5] W. P. Evans, L. Davidson, and L. Sicafuse, "Someone to listen: Increasing youth help-seeking behavior through a text-based crisis line for youth," *Journal of Community Psychology*, vol. 41, no. 4, pp. 471–487, 2013.

[6] J. Suler, "The psychology of text relationships," *Online counseling: A handbook for mental health professionals*, pp. 19–50, 2004.

[7] M. Dowling and D. Rickwood, "Online counseling and therapy for mental health problems: A systematic review of individual synchronous interventions using chat," *Journal of Technology in Human Services*, vol. 31, no. 1, pp. 1–21, 2013.

[8] R. Fukkink and J. Hermanns, "Counseling children at a helpline: chatting or calling?," *Journal of Community Psychology*, vol. 37, no. 8, pp. 939–948, 2009.

[9] M. E. Pratt, *The future of volunteers in crisis hotline work*. PhD thesis, University of Pittsburgh, 2013.

[10] K. Dinakar, J. Chen, H. Lieberman, R. Picard, and R. Filbin, "Mixed-initiative real-time topic modeling & visualization for crisis counseling," in *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 417–426, ACM, 2015.

[11] T. L. Acorn and S. H. Walden, "Smart: Support management automated reasoning technology for compaq customer service," in *Proceedings of the fourth conference on Innovative applications of artificial intelligence*, pp. 3–18, AAAI Press, 1992.

[12] R. G. Goldberg and R. R. Rosinski, "Automated natural language understanding customer service system," Apr. 20 1999. US Patent 5,895,466.

[13] R. I. Madrid, H. Van Oostendorp, and M. C. P. Melguizo, "The effects of the number of links and navigation support on cognitive load and learning with hypertext: The mediating role of reading order," *Computers in Human Behavior*, vol. 25, no. 1, pp. 66–75, 2009.

[14] M. A. Neerincx and J. Lindenberg, *Situated cognitive engineering for complex task environments*. Ashgate Publishing Limited Aldershot, 2008.

[15] J. R. Gersh, J. A. McKneely, and R. W. Remington, "Cognitive engineering: Understanding human interaction with complex systems," *Johns Hopkins APL technical digest*, vol. 26, no. 4, pp. 377–382, 2005.

[16] H. Beyer and K. Holtzblatt, *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[17] M. B. Rosson and J. M. Carroll, "Scenario based design," *Human-computer interaction. boca raton, FL*, pp. 145–162, 2009.

[18] D. Wixon, K. Holtzblatt, and S. Knox, "Contextual design: an emergent view of system design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 329–336, Citeseer, 1990.

[19] T. N. Sindahl, *Chat Counselling for Children and Youth: A Handbook*. Børns Vilkår Copenhagen, 2011.

[20] R. Tanaka, J. Itou, and J. Munemori, "Tag-based chat support system to remind users of contents of past conversations," *Procedia Computer Science*, vol. 60, pp. 891–899, 2015.

[21] W. Sunayama, Y. Shibata, and Y. Nishihara, "Continuation support of conversation by recommending next topics relating to a present topic," in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 168–172, IEEE, 2016.

[22] T. N. Nguyen and F. Ricci, "A chat-based group recommender system for tourism," in *Information and Communication Technologies in Tourism 2017*, pp. 17–30, Springer, 2017.

[23] J. Budzik and K. J. Hammond, "User interactions with everyday applications as context for just-in-time information access," 2000.

[24] J. Smithson, "Using and analysing focus groups: limitations and possibilities," *International journal of social research methodology*, vol. 3, no. 2, pp. 103–119, 2000.

[25] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 2004.

[26] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "paper recommender systems: a literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[27] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[28] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, ACM, 2008.

[29] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[30] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[31] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[32] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 238–247, 2014.

[33] R. Al-Rfou, B. Perozzi, and S. Skiena, "Polyglot: Distributed word representations for multilingual nlp," *arXiv preprint arXiv:1307.1662*, 2013.

[34] S. Tulkens, C. Emmery, and W. Daelemans, "Evaluating unsupervised dutch word embeddings as a linguistic resource," *arXiv preprint arXiv:1607.00225*, 2016.

[35] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," 2016.

[36] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, pp. 1188–1196, 2014.

[37] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International conference on machine learning*, pp. 957–966, 2015.

[38] Coosto, maintened by Alexander Nieuwenhuijse, "Dutch Word2Vec Model." https://github.com/coosto/dutch-word-embeddings. [Online; accessed November 19, 2019].

[39] J. Nielsen, *Usability engineering*. Elsevier, 1994.

[40] J. Grudin, "Utility and usability: research issues and development contexts," *Interacting with computers*, vol. 4, no. 2, pp. 209–217, 1992.

[41] W. S. Cooper, "On selecting a measure of retrieval effectiveness," *Journal of the American Society for Information Science*, vol. 24, no. 2, pp. 87–100, 1973.

[42] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[43] L. Bijker, A. Kleiboer, H. M. Riper, P. Cuijpers, and T. Donker, "A pilot randomized controlled trial of e-care for caregivers: An internet intervention for caregivers of depressed patients," *Internet interventions*, vol. 9, pp. 88–99, 2017.

[44] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[45] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

[46] K. Isbister and H. Nakanishi, "Helper agent: A chat assistant for cross-cultural conversations," *NTT REV*, vol. 12, no. 2, pp. 55–59, 2000.

Appendix A

OSF Form: Evaluation inspiration support system for suicide crisis counselling

Link to the online osf form can be found here: https://osf.io/9gu2y.

A.1 Study Information

A.1.1 Title

Experiment on the impact of a support system providing inspiration to counsellors in online suicide crisis line

A.1.2 Authors

Salim Salmi, Willem-Paul Brinkman

A.1.3 Description

Research question: How can technology provide context specific support to online chat counsellors for suicide prevention to easier execute/perform cognitive tasks? The solution this study tries to explore is giving counsellors suggestions i.e. fragments of conversations, from a text-based chat corpus, that have been matched to the current context of their ongoing conversation using an algorithm. When presented with a difficult situation the counsellors can read how their colleagues approached the situation and draw inspiration from it to more easily find their own answer. The subquestions this experiment will aim to answer are: Q1: Can experts differentiate between counsellor responses that have been provided either 1) suggestions, 2) expert feedback, or 3) no additional help? Q2: Can counsellors tell the difference between suggestions related to the context and suggestions that do not take the context into account which have been taken at random positions from the randomly chosen chats in the corpus? Q3: What is the counsellors' opinion on the usability and usefulness of the system.

Q1 will be answered by having participants read transcripts of chats that are incomplete and then providing the next message in the chat. They will do this for three chats and each time a different type of support will be given: 1) suggestions based on the context, 2) feedback from experts that have already seen the conversation, 3) no additional help Q2 will be answered by providing a transcript of a chat for the participant to read and have the participant rate suggestions provided for this transcript. The suggestions are either based on the context or random, however the participant will be blind to this. The participant must rate how much they agree that the suggestion is based on the context of the transcript they read. Experts will then blindly label which message originated from which support type. Q3 will be answered by the System Usability Scale (SUS) for measuring usability and for measuring the usefulness the participant will answer for each chat with support the question "How did the extra information help you finding your answer? " with a fixed interval scale from -3 to 3, with 3 being useful and -3 being hindering.

A.1.4 Hypotheses

H1: Experts can distinguish between responses to a difficult unfinished chat, made with help from suggestions, help from expert feedback and no help.

H2: counsellors can distinguish between suggestions found by the algorithm to be related to the context of an ongoing, unfinished chat, and suggestions that have been selected from the chat corpus at random.

A.2 Design Plan

A.2.1 Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomised controlled trials.

A.2.2 Blinding

Expert counsellors that will label the responses given by the participants will not be aware of which condition they originate from.

A.2.3 Study design

For H1 we have a within subjects design with 3 conditions (no support/suggestions/comments from experts). Each of the 6 possible ways to order the conditions will be performed an equal number of times in this experiment. Each different order of the chats will performed an equal number of times. Each participant will have a set amount of time to read a chat depending on its length and afterwards each participant has 2 minutes to formulate the next message of the

chat. After giving each reply they will give their opinion on a fixed interval scale from -3 to 3 on how the information provided for that chat has helped with finding their answer, with -3 being hindering and 3 being useful.

These replies will be labelled (blind) by experts with which condition they believe the message originated from. Before starting experiment for H1 the participant will have 5 minutes to use the support system that has the help of the suggestions to become familiar with the interface. After this the participant will fill in the System Usability Scale (SUS) for estimating usability.

For H2 we have a within subject design with 2 conditions (suggestions related to the context and suggestions randomly selected from the corpus) repeated multiple times for each participant. The participant will read 3 chats and for each chat judge 10 suggestions to be related to the context chat on a 7 point fixed interval scale with 0 being not related and 7 related. Each different order of the chats will performed an equal number of times. The condition of each suggestion will be randomised. Half are suggestions are related to the context, and half will not be related to the context and randomly selected from the chat corpus.

The pool of chats used for each hypothesis will not intersect.

A.2.4 Randomization

H1: The order of the chats used for each condition will be randomised. Which order of conditions the participant will get will be randomly assigned, but all orders will be performed an equal amount of times.

H2: This part of the experiment will involve 3 chats and 10 suggestions per chat. The order of the chats used will be randomised. The condition of each suggestion will be also randomised. The chat fragments for the second condition (not related to the chat) will be randomly selected from the chat corpus.

We assume no relationship between conditions of H1 and H2 and therefore these will be in a set order, H1 before H2.

A.3 Sampling Plan

A.3.1 Existing Data

Registration prior to creation of data

A.3.2 Data collection procedures

Recruitment will happen at 113 suicide prevention, a suicide prevention organisation in the Netherlands. Responses are collected from junior counsellors of their online chat help line.

First the participant will have a dummy conversation followed by a 5 minute window to explore and familiarise with the support system. No output is required from the participant at this time. Next three conversation transcripts that contain an unresolved difficult situation will be read by the participant. Each new line in the chat will appear after 2 seconds until the chat reaches the point where the participant has to reply. When that point is reached a timer of 2 minutes will start and they have to give their answer within that time. At this time the assisting information available for the condition will appear which the participant may use. After the participant completes the task for the condition that uses suggestions from the algorithm, they will be asked to rate on a fixed interval scale from -3 to 3 how the condition helped them in finding their answer, where 3 is useful and -3 hindering.

Afterwards the participants will fill in the SUS and the additional question for rating usability and usefulness. Finally the participants wil, for three conversation transcripts each, be presented with 10 suggestions. They will rate on a fixed interval scale of 0 to 7 if they agree with the suggestion is based on the context of the conversation they have read.

Data will be collected for a period of 2 weeks, afterwards 4-6 experts will blindly label the gathered responses with what condition they believe the response was made with.

A.3.3 Sample size

24 junior counsellors.

A.3.4 Sample size rationale

24 participants is an amount that is achievable in the available time with the amount of counsellors available. We decided on 24 participants as it is easily divisible by the amount of different orders of our conditions for H1.

A.3.5 Stopping rule

Data collection will be terminated when 24 complete data sets have been collected (no missing elements).

A.4 Variables

A.4.1 Manipulated variables

H1: Three levels nominal: Whether the counsellor receives no help, help from the algorithm or help from the expert.

H2: Two levels nominal: Whether the suggestion presented to the participant is a suggestion found by the algorithm based on the context of the chat or if the suggestion is randomly selected from the chat corpus without looking at the context.

A.4.2 Measured variables

H1:

- The textual response the counsellor gives.

- The labels the experts give to the responses of the counsellor (no support/ suggestion from algorithm / expert help)

H2: The rating the counsellor gives for the suggestion a suggestion on a 7 point fixed interval scale on if the counsellor agrees the suggestion is related to the context of the chat. The chat the suggestion was given for.

The System Usability Scale (SUS).

Usefulness score: Rating on a 7 point fixed interval scale from -3 to 3

A.4.3 Indices

SUS combine answer into a score using the mean and standard deviation.

A.5 Analysis Plan

A.5.1 Statistical models

H1: We will use a multi-level mixed model with a crossed random effect. The manipulated nominal independent variable is the type of support (no support/ suggestion from algorithm / expert help) the participant received. The dependent variable the label given by the expert. Participant and expert will be included as random effects with a variable slope and intercept.

H2: We will use a multi-level mixed model. The independent variables will be the participant the chat and the suggestion (fragment of a conversation that is related to the context or not related to the context). The dependent variable is the rating the participant gives for how much they agree that the suggestion is based on the context.

For the usefulness score we will use a one sample t-test compared with 0.

Confidence interval for SUS

A.5.2 Transformations

SUS responses of the participants will be normalised.

A.5.3 Inference criteria

H1: Post-hoc comparison between three conditions, with alpha = 0.016

H2: P-values smaller than 0.05

SUS: P-values smaller than 0.05

Usefulness score: P-values smaller than 0.05

A.5.4 Missing data

Data samples that are incomplete (unfinished or not completed in time) will be excluded.

Appendix B

Multi-level models

Figure B.1 shows the simple 1 level mixed model, used in Section 5.2.6. The conditions are denoted with c and the participating counsellors with P. Figure B.2 illustrated the mixed model with a crossed random effect, with counsellors P and experts E, used in Section 5.2.6.

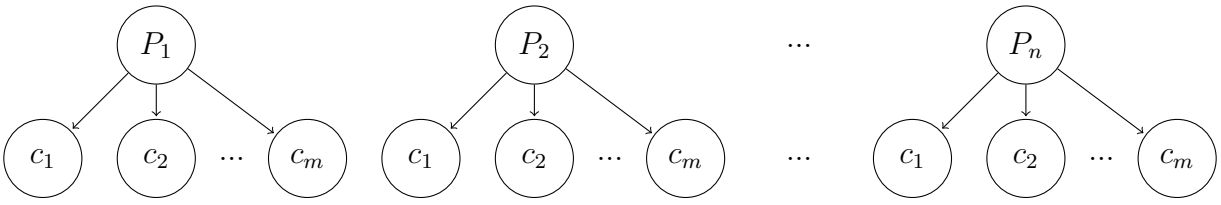


Figure B.1: Linear mixed effect model

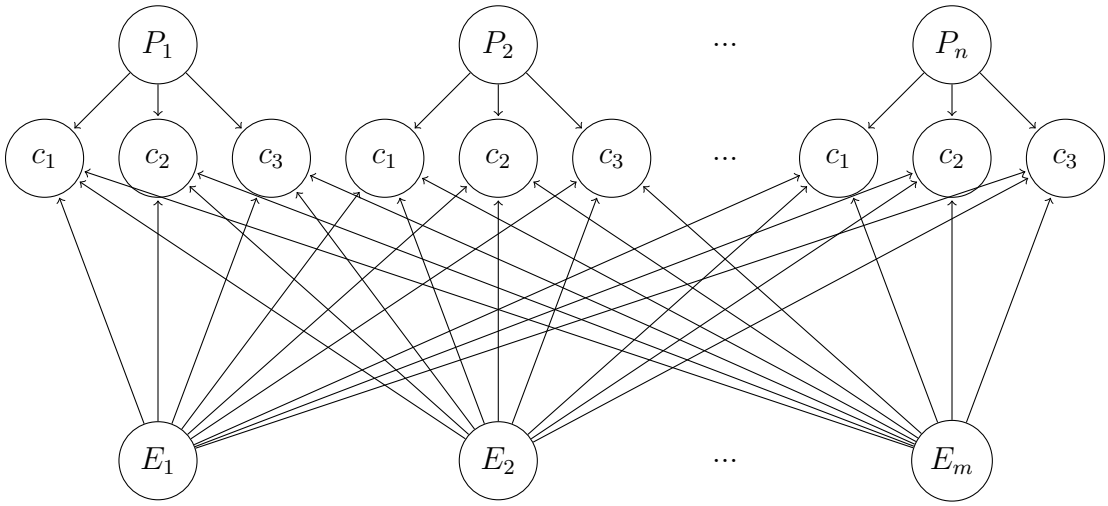


Figure B.2: Crossed random effect model

Appendix C

Assumptions plots

This appendix contains the plots for confirming the assumptions the mixed model logistic regression described in Section 5.2.6. As well as the results using the log-transformed data, as described in Section

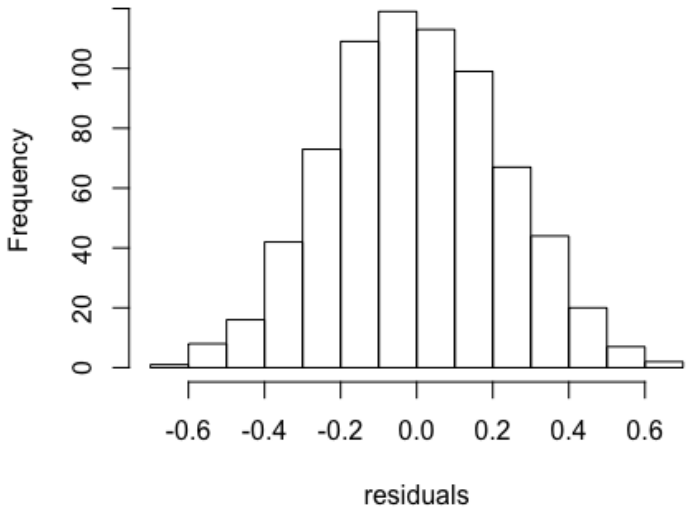


Figure C.1: Histogram of the residuals of mixed model; performance of the algorithm

The results for unaltered counsellor showed that the suggestion type significantly predicted counsellor rating b=1.07, t=7.66, p< .0001. The results for log-transformed data showed

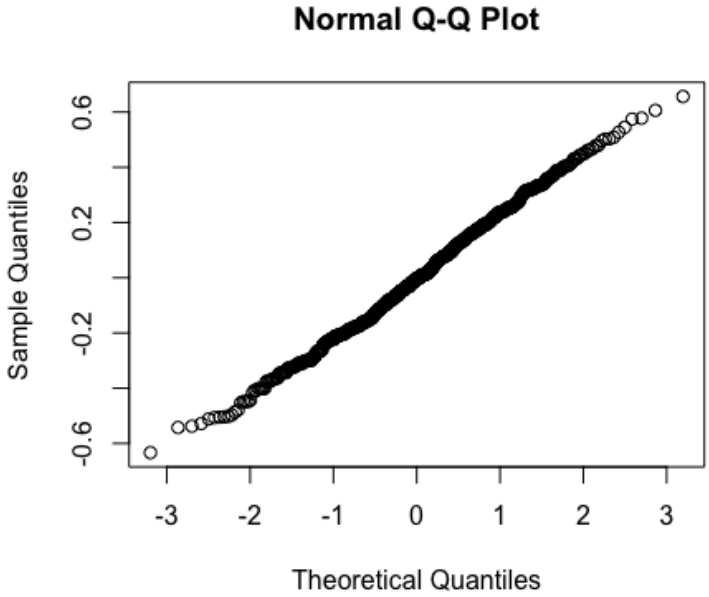


Figure C.2: Q-Q plot of mixed model; performance of the algorithm

suggestion type significantly predicted counsellor rating, b=0.40, t=10.08, p< .0001.

Model comparison	χ^2(degrees of freedom)	P value
Log-transformed measures		
Add suggestion type	94.55(1)	<.0001

Table C.1: Model comparisons for effects of suggestion type and chat on counsellor rating (n=720)

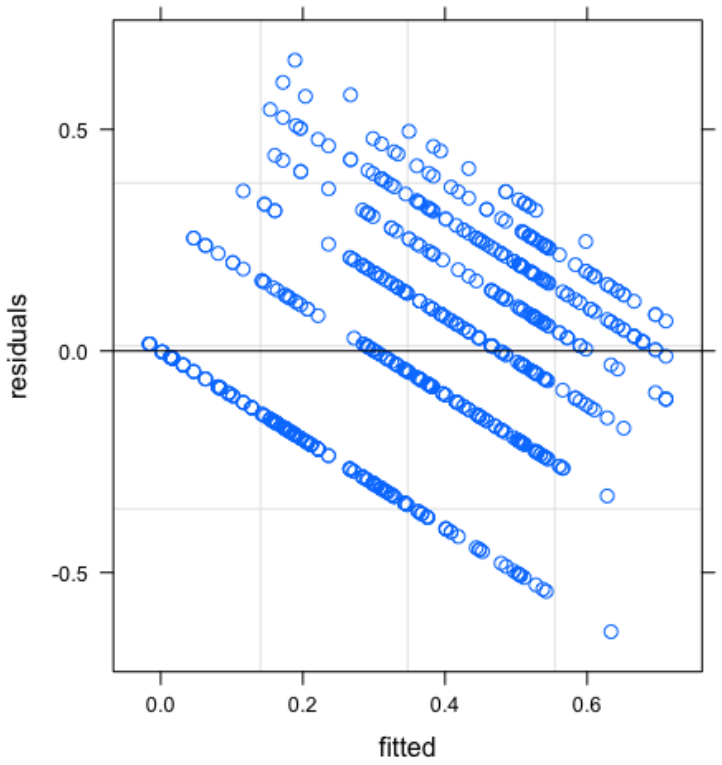


Figure C.3: Plot of residuals and fitted values of mixed model; performance of the algorithm

Appendix D

Consent form

This document is to inform the participants of the research performed by TU Delft student Salim Salmi supervised by dr. Saskia Mérelle (113) and dr.ir. Willem-Paul Brinkman (TU Delft).

The purpose of this research is to evaluate a potential support system aimed at giving inspiration to during a moment where a counsellor might find it difficult to make progress with the caller and the counsellor has trouble finding an appropriate response to move the conversation forward.

The benefits of this research will include understanding the effectiveness of generated suggestions for crisis counselling and counsellor satisfaction with the potential support system.

In this evaluation experiment you will be presented with 3 transcripts of a chat that is unfinished. The goal for the participant of this study is to pick up where the chat left of and write the next message. The participant can be assisted in one of three ways (picked at random):

- Conventional 113 macros
- Conventional 113 macros and advice from a floor manager
- Conventional 113 macros and a support system

This study will record the participant's message as well as how satisfied the participant is with the message and how satisfied the participant is with the work setting they were assigned. Following this experiment the participant's answers will be blindly categorized by psychologist at 113.

Following this the participant is presented with an additional 3 transcripts. For each of these transcripts several suggestions by the support system are given. The participant will be asked which of these suggestions fit the context of the conversation presented in the transcript.

The risks involved in this study are the identification of the counsellor or trainers and that any answer they may give during the experiment be used to judge their performance. To prevent this, during this study no individual identifiable data will be collected as the data will be store with an anonymized code. The result of the experiment will only be available to the relevant researchers as named above. The messages will be read by psychologist but without any other identifiable information. The raw data will be stored for a maximum period of 9 months, the collected results will be anonymized, and only summaries of the study will be published. Note that during the study the participant can withdraw at any moment.

I have read and understood the study information dated 23/05/2019, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.
I understand that taking part in the study involves reading an artificial chat transcript of a conversation on the 113 chat platform with a person that seeks help regarding suicide and typing a response. I understand this response will be stored without any identifiable information, and will be read by 113 psychologists.
I understand that taking part in the study involves the following risks: Identification, which will be minimized by anonymizing the data.

I understand that information I provide will be used for evaluation purposes of a potential support system for crisis counselling and the results might be used in an anonymized format for scientific publications. The anonymized and summarized results will be included in a corresponding report.

I understand that personal information collected about me that can identify me, such as my name will not be shared beyond the study team.

I agree that my information can be quoted in research outputs

I agree to joint copyright of the reactions, questionnaire and ratings to Salim Salmi.

I give permission for the reactions, questionnaire and ratings that I provide to be archived in TU Delft repository so it can be used for future research and learning.