From Words to Actions

Natural Language Processing for suicide prevention helplines

July 12, 2024

Dissertation Committee

promotors: prof.dr. R.D. van der Mei
Centrum Wiskunde & Informatica, Amsterdam, the Netherlands
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

prof.dr. S. Bhulai
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

copromotors: prof.dr. R. Gilissen
113 Zelfmoordpreventie, Amsterdam, the Netherlands
Universiteit Leiden, Leiden, the Netherlands

dr. S.Y.M. Mérelle
113 Zelfmoordpreventie, Amsterdam, the Netherlands

committee: prof.dr. A. Fokkens (voorzitter)
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

prof.dr. A. Yaman
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

prof.dr. Mishara
University of Québec at Montréal, Montréal, Canada
Centre for Research and Intervention on Suicide, Ethical Issues and End-of-Life Practices, Montréal, Canada

prof.dr. A.T.F. Beekman
Amsterdam UMC, Amsterdam, the Netherlands

dr. L. Schweren
113 Zelfmoordpreventie, Amsterdam, the Netherlands

Vrije Universiteit

From Words to Actions

Natural Language Processing for suicide prevention helplines

Academisch Proefschrift

ter verkrijgen van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus,
in het openbaar te verdedigen
ten overstaan de promotiecommissie
van de Faculteit der Bèta wetenschappen
op ...
in een bijkomst van de universiteit,
De Boelelaan 1105

door

Salim Salmi

geboren te Rotterdam

List of Abbreviations

NLP	Natural Language Processing
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
SBERT	Sentence-BERT
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
LLM	Large Language Model
MI	Motivational Interviewing
LDA	Latent Dirichlet Allocation
LIME	Local Interpretable Model-agnostic Explanations
BoW	Bag-of-Words

Contents

Part II Topic modeling for helplines

Part III Classification and insights

Chapter 1

General introduction

1.1 A public health concern

In the field of public health, few issues command as much urgency and compassion as the global crisis of suicide. Defined as the act of intentionally causing one's own death, suicide represents a profound challenge to societies around the world, transcending geographical, cultural, and socioeconomic boundaries. With the World Health Organization (WHO) estimating that approximately 700,000 lives are lost to suicide annually [181], the severity of this issue cannot be overstated. Behind these stark statistics lie immeasurable pain, shattered families, and communities left grappling with the aftermath of profound loss.

While suicide is undoubtedly a multifaceted phenomenon influenced by complex factors, it is crucial to acknowledge the pivotal role that helplines play in addressing and mitigating this public health concern. Suicide prevention helplines serve as lifelines, offering support and understanding to individuals in distress [141].

1.2 Suicide prevention helplines

Suicide prevention helplines are specialized services designed to offer immediate support, intervention, and resources to individuals experiencing suicidal thoughts. These helplines are a preventive service to reduce the suicidal ideation or behavior of help seekers [22]. Helplines play a crucial role in providing a confidential and non-judgmental space for individuals to share their feelings, thoughts, and concerns, with the ultimate aim of preventing suicide.

In the current landscape, suicide prevention helplines have evolved to include online platforms, where individuals can access support through various digital channels such as mobile phone calls, text messages, or online chat services. These online helplines often employ trained and dedicated volunteers and counselors who offer their time and expertise to engage with those in need. They undergo specialized training to equip them with the skills necessary for handling crisis situations [2].

Some helplines provide anonymity as a cornerstone of their service, allowing individuals to seek assistance without the fear of judgment or disclosure of their identity. The confidentiality provided by suicide prevention helplines is crucial in encouraging open communication and fostering a safe environment for individuals to express their struggles.

In addition to providing immediate emotional support, suicide prevention helplines often serve as gateways to additional resources, such as mental health professionals, community services, or crisis intervention teams, ensuring that individuals receive the comprehensive assistance they require.

1.3 The current state of helpline research: What are the challenges and priorities?

Despite their importance, helpline research faces several challenges. The current landscape demands that they prioritize certain areas to enhance their effectiveness and continue their operations with a demand that is still growing. This section summarizes some of the challenges that motivate the research in this dissertation.

The taxation of counseling: Counseling is an emotionally demanding and psychologically taxing profession [48]. The weight of the responsibility to support individuals in crisis, coupled with the inherent difficulty of the subject matter, can contribute to high attrition rates among helpline counselors. A resilient and consistant body of counselors is important to maintain a helpline's increasing demand.

Operational awareness: Helpline managers must be able to stay informed about the functioning of the service. This includes monitoring conversations, and address any emerging themes.

Enhancing quality of care: Suicide prevention helplines are a lifeline for many, and as such, there is an inherent responsibility to continually improve the quality of care provided. However, anonymity is a cornerstone of many suicide prevention helplines, who hope it fosters an environment where individuals feel safe to share their struggles. It is therefore not feasible to perform large longitudinal studies to evaluate and experiment as a means of improving the helpline.

Insight into the helpline population: Gaining insight into the challenges faced by individuals contemplating suicide is crucial for effective helpline support. This information helps in tailoring support and connecting help seekers with appropriate resources, ensuring timely intervention and support. It also supports the field of suicide research overall, by gaining insights on a large at-risk population.

1.4 NLP as a tool to address helpline difficulties.

Natural Language Processing (NLP) could provide an avenue to addressing these difficulties. NLP has been a powerful tool to tackle problems in many fields, including mental health and suicide prevention.

1.4.1 What is NLP?

NLP is a subfield of Artificial Intelligence (AI) that focuses on the interaction between computers and human language. It involves the development of algorithms and models that enable computers to understand, interpret, and generate human-like text.

The concept of automated text analysis has existed in some form for a very long time [1]. Basic rule-based occurrence of a set of words is still a powerful method given the right domain and goal. However, NLP as a field started gaining traction with statistical modeling. Techniques like topic modeling started to produce good results. Latent Dirichlet Allocation (LDA) [14], among others, is a powerful method to uncover latent thematic structures in a set of documents.

These approaches are often dubbed Bag-of-Words (BoW) method, represent documents as unordered collections of words. However, this can have some significant drawbacks. The same word in two different sentences can have two different meanings, depending on the context, but BoW methods make the assumption that this is not the case. Language is often structured to be contextual and there are numerous words that can have different meanings. For example, depending on the context, the word "grand" can mean large, magnificent or a thousand.

Taking the context of the sentence into account is easier said than done. The following is a classic example that illustrates the potential difficulty. For example, consider the two sentences:

1. *The trophy would not fit in the suitcase because it was too big.*
2. *The trophy would not fit in the suitcase because it was too small.*

Humans can intuitively understand that in sentence 1, the word *it* refers to the trophy being too big, whereas in sentence 2, the word *it* refers to the suitcase being too small. Now suppose you want to translate this sentence into a language like French, where trophy and suitcase have different genders, *it* would need to be translated differently. A computer would need to be able to understand this spacial relationship.

To be able to perform contextual tasks like this, Machine Learning (ML) approaches have become a prevalent tool in NLP [142]. ML is a subset of AI that involves the development of algorithms and models that enable computers to learn patterns and make predictions or decisions without being explicitly programmed. It relies on the use of data to allow systems to improve their performance on a specific task over time through experience. To perform contextual tasks, the models that are most often used are Artificial Neural Networks (ANN) [5]. ANNs are computational models, often compared to the structure and functioning of the human brain, consist of interconnected nodes (neurons) that are organized into layers, including an input layer, one or more hidden layers, and an output layer. ANNs are designed to learn from data, adjusting the weights of connections between neurons to make predictions, classify information, or perform various tasks through iterative training processes.

1.4.2 Transformers

The Transformer architecture, introduced by Vaswani et al. [180] in the paper "Attention is All You Need", revolutionized NLP and various other ML tasks. Before Transformers, Recurrent Neural Networks (RNN) [3] and Long Short-Term Memory networks (LSTM) [8] were commonly used ANNs for sequence-to-sequence tasks [49]. However, these models had limitations in handling long-range dependencies and parallelization.

The Transformer architecture is an ANN and consists of an encoder-decoder structure, each containing multiple layers. The key innovation lies in the attention mechanism, which allows the model to weigh different parts of the input sequence differently during processing. The attention mechanism is a crucial component of the Transformer architecture. It enables the model to focus on specific parts of the input sequence when making predictions for a given token. The attention mechanism can be defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

Here, Q, K, and V represent the query, key, and value weight matrices, respectively. The scaled dot-product attention uses the softmax function to normalize the attention weights, and d_k is a scaling factor. During training these three matrices are what capture the different aspects of the tokens.

The attention operation can be seen as akin to a database search operation. The Q "Query" matrix could be seen as a representation of what is being looked for. The K "Key" matrix could be seen as a representation of the context against which the Q matrix searches. The V "Value" matrix represents the information that is relevant from each token. The resulting attention operation is a weighted average of the important information, of each token, for each token.

The Transformer architecture applies a version of attention called *multi-head* attention. Each attention layer performs the attention task multiple times, with seperate weights, called *heads*. The rationale being that each head learns different aspects. The resulting states are subsequently projected down to the original dimension for the next layer.

The attention mechanism has two key strenghts that has allowed it to nearly universally become the best performing ANN architecture in language representation and language generation tasks.

1. **Capturing long-range dependencies:** Attention allows the model to capture dependencies between different parts of the input sequence, enabling it to consider relevant information when making predictions. Transformers excel at capturing long-range dependencies, making them suitable for tasks that require understanding relationships between distant tokens.

2. **Parallelization:** Unlike sequential processing in RNNs, Transformers can process tokens of an input sequences in parallel, leading to more optimized training.

The Transformer architecture quickly became the state-of-the-art in various NLP tasks and beyond due to these scalability, parallelization capabilities, and ability to capture long-range dependencies. Models derived from Transformers have achieved top performance in tasks such as *machine translation*, text summarization, question answering, and more. These Transformer-based models have significantly advanced the field of natural language processing and have been adapted for various domains, demonstrating the transformative impact of the attention mechanism and the Transformer architecture.

The main application for the Transformer was machine translation. The encoder creates a representation of the text to be translated.

The decoder generates the translation token by token, based on the representation and the tokens that it had generated thus far.

1.4.3 Pretrained Transformers and transfer learning

Looking back to the problem of the trophy and the suitcase, not only does one need to be able to process a sentence contextually, to relate *it* to different nouns, one also needs to have some knowledge of the world to know what trophies and suitcases are and how they relate to each other. We would need a large amount of data if a model needs to learn these things for every different NLP task. That is why pretrained Transformers have become the most common way to train models in NLP for many tasks.

Pretrained Transformers refer to models that are trained on massive datasets for a general NLP task, such as token prediction. Transfer learning involves leveraging these pretrained models for downstream tasks with limited labeled data, the idea being that all the general concepts of language are already learned, and we only need to fine-tune it to specialise it for new tasks. By fine-tuning the pretrained models on specific tasks, they can quickly adapt to new domains and achieve good performance with less labeled data.

Two models stand out in this regard as the most popular methods of pre-training based on the Transformer architecture:

- **Bidirectional Encoder Representations from Transformers (BERT) [98]:** Introduced by Google in 2018, BERT pretrains a Transformer model on a large corpus and achieves impressive results on various downstream NLP tasks. BERT is based on the encoder part of the traditional Transformer and is most often used to create representations of text or text classification.

- **Generative Pre-trained Transformer (GPT) [92]:** Developed by OpenAI, GPT is a series of models that use Transformers for language understanding and generation. GPT-4, in particular, is one of the largest language models to date. GPT is based on the decoder part of the traditional Transformer and is mainly used for text generation.

A common that can be performed with BERT is text embedding. Text embedding refers to the process of representing words or documents as vectors in a continuous vector space. This representation captures semantic relationships between words, enabling algorithms to understand the contextual meaning of words. Word embeddings, such as Word2Vec, GloVe, and FastText, are widely used for this purpose.

Another common task performed with BERT is classification. Classification is a fundamental task in NLP where the goal is to categorize input text into a predefined set of classes or labels. It involves training a model on a labeled dataset, allowing it to learn patterns and associations between the input text and corresponding classes. For example, a common classification tasks is to predict whether an email is spam or not.

1.5 Possible avenues to include NLP technology in helplines

In the realm of suicide prevention helplines, integrating NLP technology opens up promising avenues for leveraging existing data, collecting additional information, and testing and evaluating tools to support their operation. The following are the three main avenues in this dissertation to include NLP in helplines.

Leveraging existing data: Suicide prevention helplines naturally generate a vast amount of textual data through phone calls, chat conversations, and other interactions. NLP can be applied to analyze this wealth of information retrospectively. By processing transcripts and call logs, NLP algorithms can identify patterns, detect common themes, and extract valuable insights regarding the experiences of both help seekers and counselors. This retrospective analysis can inform improvements in training programs, identify areas for intervention, and enhance the overall effectiveness of the helpline.

Large-scale data collection through questionnaires: NLP technology can be employed predict and analyse outcomes from questionnaires. These questionnaires can be integrated into the online chat platform, allowing helplines to collect structured data on a wide range of topics related to mental health, crisis situations, and user satisfaction. NLP algorithms can then analyze the responses, extracting meaningful information and possibly make connections between the demographics, needs, and concerns of helpline users. This data-driven approach facilitates evidence-based decision-making, enabling helplines to adapt and evolve in response to the dynamic landscape of mental health challenges.

Testing and evaluation of helpline tools: NLP can play a pivotal role in testing and evaluating the efficacy of various helpline tools and interventions. For instance, during chats NLP analysis can provide information based on the current contents of the conversation. The chat outcomes can gauge effectiveness, helping helplines fine-tune these tools for optimal performance. This iterative process of testing and evaluation, ensures that helpline interventions are evidence-based, user-friendly, and aligned with the evolving needs of those seeking assistance.

By integrating NLP technology into these areas, suicide prevention helplines can harness the power of existing data to enhance their service and inform future approaches. This data-driven approach empowers helplines to tailor interventions to individual needs, and continuously improve their effectiveness in their mission of preventing suicide.

1.6 113 Suicide Prevention

The work in this dissertation is primarily conducted in collaboration with the national Dutch suicide prevention helpline, "113 Suicide Prevention" [184]. The main goal of 113 is to reduce the number of suicides and suicide attempts by providing accessible, professional, and effective support services. This helpline aims to offer immediate

Table 1.1: Helpline questionnaire

Label	Options
Name	a name or alias of the help seeker
Age	the age of the help seeker
Gender	man, woman or other
Living situation	alone, with a partner, with children, with parents, in a psychiatric hospital, with housemates, other
Treatment	if the help seeker is receiving treatment at the moment, yes or no
Work or School	School, employed, self-employed, unemployed, incapacitated, homemaker, retired, other

assistance to those in crisis and work towards a society where discussing mental health and seeking help is normalized.

113 operates a free, anonymous helpline available 24 hours a day. Help seekers can contact trained counselors through either a telephone service or a chat service accessible via their website. In 2023 there were a total of 184.515 conversations in the helpline with 52% chat conversations and 48% telephone conversations [185]. The data from the chat service, which is more structured, serves as the primary source for the research in this dissertation.

Below you find an overview of the process: help seekers first access the 113 website. From the home page, they can enter the chat service. Before being placed in a queue, they answer an automated survey with several questions. Table 1.1 lists all the basic questions asked. Afterward, the help seeker is asked to rate themselves on a scale of 1 to 7 on several risk factor questions. Table 1.2 presents these questions and the specific items assessed.

After completing the questionnaire, help seekers are placed in a queue. One or more reception counselors then accept their chat request. At this stage, the counselor performs a triage by asking about the reason for the chat, assessing the help seeker's safety, and then forwarding the conversation to another counselor for further discussion. This triage focuses on a limited set of questions and often involves handling multiple help seekers simultaneously.

During the actual helpline conversation, the counselor operates

Table 1.2: Suicide risk factor items

Variable	Item
Suicidal ideation	I feel the urge to kill myself
Unbearable psychache	I can't take my pain anymore
Hopelessness	I feel hopeless
Defeat	I feel that I have given up
Entrapment	I feel trapped
Perceived burdensomeness	I am a burden to others
Thwarted belongingness	I feel like I do not belong
Desire to live	I have the desire to live
Capability for suicide	I could kill myself if I wanted to

according to the Motivational Interviewing (MI) paradigm. This approach is a client-centered, directive method for enhancing intrinsic motivation to change by exploring and resolving ambivalence. Counselors use open-ended questions, reflective listening, and affirmations to help individuals articulate their goals and motivations for change.

Finally, after the conversation, the help seeker is requested to fill in an optional post-chat questionnaire, with the same items as in Table 1.2.

The work in this dissertation is enabled through the text and questionnaire data that is gathered from these chat conversations.

1.7 Ethics and bias

Incorporating NLP technologies into suicide prevention helplines raises critical ethical considerations and the challenge of mitigating bias. The sensitive nature of helpline communications necessitates a careful approach to ensure these technologies enhance, rather than compromise, the quality of support provided. NLP applications must rigorously protect the privacy and confidentiality of individuals, adhering to stringent data protection standards. Therefore, all data is properly anonymized before used for model training and research.

Another important ethical consideration is the need for human contact for help seekers in crisis. It is crucial to ensure that help seekers do not feel like they are interacting with a robot, as genuine human interaction is essential. This contrasts with customer support services, where customers are typically seeking practical solutions to specific problems. In a helpline context, help seekers are often dealing with profound and often urgent emotional distress, requiring empathy, understanding, and emotional support. Even if an AI system can replicate this interaction, help seekers expect to be communicating with a human, and if this expectation is not met, it constitutes a breach of trust.

1.8 Research questions

This dissertation seeks to explore the potential of NLP within the context of suicide prevention helplines. The research is guided by two fundamental questions.

1. *How can NLP techniques support suicide prevention helplines?* The first question delves into the practical applications of NLP in supporting the operations and staff of suicide prevention helplines. The goal is to identify NLP-driven strategies that can alleviate the burdens of counselors, and enhance the awareness of current problems within the help seeker population, and ultimately, contribute to more effective suicide prevention efforts.

2. *What new insights can be gained from helpline data with NLP Techniques?* The second question focuses on the analytical power of NLP to extract meaningful patterns and insights from the extensive data collected by suicide prevention helplines. Through the analysis of conversation logs, this research aims to uncover trends, needs, and potential areas for intervention that are not immediately apparent. This question seeks to harness NLP's ability to process and analyze large datasets to provide a deeper understanding of the dynamics at play in suicide prevention interactions, offering a foundation for evidence-based improvements and innovations in helpline

services.

Together, these research questions aim to bridge the gap between technological advancements in NLP and the practical, on-the-ground needs of suicide prevention helplines. By answering these questions, this dissertation hopes to contribute to a future where technology and compassion converge to offer stronger support to those with suicidal thoughts.

1.9 Overview of this dissertation: Three levels of helpline assistance with NLP

This dissertation is divided into three parts. Each part aims to address challenges on a different level in the helpline.

1.9.1 Part I: Support system for counselors through deep learning recommendation

The field of suicide prevention counseling is intensely demanding, sometimes resulting in counselor burnout due to its significant emotional toll. The maintenance of counselor well-being and their ongoing professional growth is crucial to sustaining high-quality support services. Although counselors receive training in critical conversational techniques, such as MI, research indicates that these skills can diminish over time without continuous practice and feedback. Furthermore, it is important to maintain quality in helpline conversations, especially during challenging situations.

To address this challenge, we propose the adoption of a a deep learning recommendation system. This system analyzes patterns and nuances in past counseling sessions, identifying successful strategies and areas for improvement. Counselors receive personalized, real-time feedback, allowing them to enhance their skills and adapt their approach. This continuous feedback loop not only supports counselors in their professional development but also serves as a preventive measure against burnout or *compassion fatigue*.

Chapter 2 of this dissertation outlines the development and preliminary testing of this promising support tool. Subsequently, Chapter 3 delves deeper, focusing on enhancing the algorithm's effectiveness and rigorously evaluating its impact through a real-time randomized control trial. This research aims to underscore the potential of deep learning recommendations in revolutionizing counselor support mechanisms, thereby improving the quality of care offered to those in crisis.

1.9.2 Part II: Topic modeling for day-to-day operation of the helpline

Suicide prevention helplines handle a diverse range of topics and issues, making it crucial for the helpline managers to gather operational insights to allocate resources effectively and understand the demographic. The challenge lies in understanding the backgrounds, problems and evolving needs of help seekers such that the helpline can provide the necessary support.

Implementing topic modeling algorithms on chat transcripts provides a solution to this challenge. By extracting prevalent themes and issues, helpline managers gain a nuanced understanding of the day-to-day functioning of the helpline. This insight can facilitate the development of tailored responses to current issues or the provision of additional information on specific topics. The result is an adaptive and responsive helpline that can efficiently address the evolving needs of its users.

Chapter 4 delves into the exploration of topic modeling techniques, with a focus on their application to counseling dialogues. Chapter 5 implements these findings, by applying the best methodologies to conversations captured before and during the initial COVID-19 lockdown in the Netherlands. This sequential analysis aims to demonstrate how topic modeling can not only enhance operational efficiency but also ensure the helpline remains responsive and relevant amidst crisis situations.

1.9.3 Part III: Enhancing understanding and functionality with deep learning

The search to gauge and augment the efficacy of suicide prevention helplines is hindered by the impracticality of longitudinal studies, which often compromise privacy. While small-scale analyses have been performed, the exhaustive coding required for broader datasets is not feasible through manual efforts.

Deep learning, particularly through the use of Transformer models like BERT, offers a compelling solution, as they have significantly boosted classification performance in recent years. This advancement has made it possible to classify vast quantities of messages swiftly, opening new avenues for understanding and improving helpline operations without infringing on privacy.

Classification is an important step towards learning to recognise patterns for specific outcomes. In Chapter 6, we delve into classification approaches to classify patterns of MI within helpline interactions. The classification effort is continued in Chapter 7, which explains a hierarchical Transformer model aimed at categorizing helpline conversations according to their outcomes. This model was then interpreted to identify key messages that influence classification outcomes, yielding insights for enhancing helpline strategies and support.

In summary, these detailed approaches utilize deep learning to provide a supportive environment for counselors, enhance day-to-day operational efficiency, and holistically assess and improve the overall functioning of suicide prevention helplines.

1.10 Publications of the author

This thesis is based on the following publications:

- Chapter 2 is based on [144]: Salmi, S., Mérelle, S. Y. M., Gilissen, R. & Brinkman, W. P. Content-Based Recommender Support System for Counselors in a Suicide Prevention Chat

Helpline: Design and Evaluation Study. *Journal of Medical Internet Research* **23,** e21690 (Jan. 2021)

- Chapter 3 is based on [190]: Salmi, S., Mérelle, S. Y. M., van Eijk, N., Gilissen, R., van der Mei, R. D. & Bhulai, S. Real-time assistance in suicide prevention helplines using a deep learning-based recommender system: a randomized controlled trial. Submitted for publication.
- Chapter 4 is based on [186]: Salmi, S., van der Mei, R. D., Mérelle, S. Y. M. & Bhulai, S. Topic modeling for conversations for mental health helplines with utterance embedding. *Telematics and Informatics Reports* **13,** 100126 (Mar. 2024)
- Chapter 5 is based on [166]: Salmi, S., Mérelle, S. Y. M., Gilissen, R., van der Mei, R. D. & Bhulai, S. Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID-19 pandemic: in-depth analysis using encoder representations from transformers. *BMC Public Health* **22** (Mar. 2022)
- Chapter 6 is based on [188]: Pellemans, M., Salmi, S., Mérelle, S. Y. M., Janssen, W. C. & van der Mei, R. D. Automated Behavioral Coding to Enhance the Effectiveness of Motivational Interviewing in a Chat-Based Suicide Prevention Helpline. *Journal of Medical Internet Research.* Accepted for publication.
- Chapter 7 is based on [189]: Salmi, S., Mérelle, S., Gilissen, R., van der Mei, R. D. & Bhulai, S. The most effective interventions during online suicide prevention chats: Machine Learning Study. Under review.

Part I

Assisting counselors

Chapter 2

Designing a support tool for a suicide prevention helpline

Salmi, S., Mérelle, S. Y. M., Gilissen, R. & Brinkman, W. P.

Journal of Medical Internet Research, **23** e21690 (Jan. 2021)

Abstract

Background

The working environment of a suicide prevention helpline requires high emotional and cognitive awareness from chat counselors. Counselors believe that as chat conversations become more difficult, it takes longer to compose a response, often leading to writer's block. This study evaluates and designs supportive technology to help counselors overcome writer's block in challenging chat situations.

Methods

A content-based recommender system using sentence embedding was developed to search a chat corpus for similar situations. The system showed counselors the most relevant parts of former chats for inspiration. In a within-subject experiment, counselors' replies were analyzed under three conditions: using the support system, receiving advice from a senior counselor, or receiving no help. The system's utility, usability, and algorithm validity were also assessed.

Results

Twenty-four counselors tested the prototype. Experts could significantly predict if counselors received help from the support system or a senior counselor ($p = .004$). Counselors found senior counselor advice ($M = 1.46, SD = 1.91$) significantly more helpful than the support system or no help ($M = -0.21, SD = 2.26$). Additionally, counselors rated chat conversations identified by the recommender system as significantly more similar to their current chats ($\beta = 0.30, p < .001$).

Conclusion

Support influenced counselors' responses in difficult conversations. However, higher utility scores for senior counselor advice suggest that specific, actionable instructions are preferred. These findings will aid in developing a system that generates advice from similar chat situations, helping counselors overcome writer's block.

2.1 Introduction

Historically, people have been able to contact helplines over the telephone, but with the advent of the internet, chat services have become increasingly popular. Compared to telephone helplines, online chat helplines show approximately the same beneficial effects [63]. Help seekers mention several reasons for using counseling through an online chat rather than a traditional phone call, such as privacy and the slow deliberate nature of online chatting [27, 35, 39, 40]. In the Netherlands, 113 saw the number of conversations increase to more than 35,000 via telephone and more than 57,000 via online chat in 2018, an increase of 33% from 2017. However, this increase resulted in a higher need for counselors as well. Because of the difficult nature of crisis counseling, suicide prevention helplines often have difficulty retaining counselors [43].

Studies have indicated that technology can support chat line operators in executing cognitive tasks. For example, in the related field of commercial telephone and chat customer support, there are various supportive technologies developed for operators [4, 9, 28]. However, in computing research aimed at suicide prevention, most work focuses on the prediction and detection of suicidal behavior [82, 83], while only a few studies have examined assisting online counselors; this could be beneficial, though. Salmi [109] has identified several difficulties that counselors encounter in their work. First, the counselor has to take in a large amount of information about the help seeker. Here, counselors could be supported in understanding a help seeker's history without having to read large portions of transcripts. Dinakar et al. [52], therefore, have created a support system prototype for text-based crisis counseling called Fathom. Fathom uses visualizations based on topic modeling to provide information at a glance. In comparison to a control interface without a visualization aspect, Fathom was preferred by counselors when eliciting a list of issues and a conversation summary. Another difficulty is that the counselor must be aware of the conversation quality. In this respect, Althoff et al. [62] compared the chat conversations of more and less successful counselors with natural language processing techniques to discover the quality differences, defining actionable strategies to improve conversation quality. For example, they showed that more successful counselors spend a longer time exploring solutions, while less successful counselors spend more time defining the problems.

Finally, the complexity and severity of help seekers' situations may lead to writer's block in counselors. Although not directly related to the suicide prevention domain, Isbister et al. [10] have designed a helper agent for human-human interaction. When a conversation lags, the agent suggests topics for the conversation pair to talk about and, thereby, the agent is generally able to make positive contributions to the chat.

In situations where counselors experience writer's block, a straightforward solution would be to approach a senior colleague for help. These senior counselors can read along and describe in as much detail as necessary how they would respond to the help seeker. However, this requires availability and time from a colleague, and this is not always possible. Responding quickly is important in life-threatening situations, and counselors cannot always wait for somebody to become available. We also suspect that an approach such as suggesting topics to keep a conversation going [165] or providing a conversation summary [83] would not be optimal in difficult situations where counselors have to de-escalate a suicide-related crisis. This paper, therefore, presents a system that uses natural language processing techniques to provide support for counselors in difficult chat conversations. The system recommends parts of similar, previous chat situations for the counselor to draw inspiration from, which might be able to reduce their writer's block. This paper also evaluates the designed support system by comparing it with 1) written, general advice from a senior counselor and 2) receiving no additional help during chats. The system's usability and utility, along with the validity of the algorithm used, were also examined.

2.2 Methods

2.2.1 Design

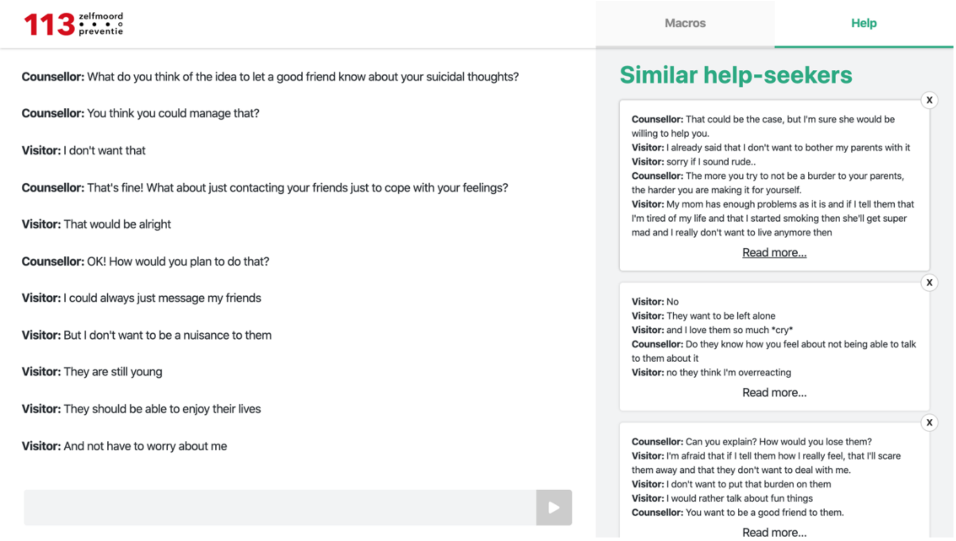
We used a within-subject design to evaluate the impact and usefulness of similar chat situations that could be used as inspiration. In the study, the counselor wrote a chat reply to a simulated chat that was interrupted as a difficult situation. The counselor took part in three simulations: 1) the counselor received parts of similar chats from a support system, 2) the counselor received written advice from an experienced counselor, and 3) the counselor received no additional help. A questionnaire was used to measure the support system's usability. Finally, we evaluated the validity of the similarity of the generated chats by testing the algorithm in a small additional experiment with a within-subject design.

The current study received ethical approval from the TU Delft University research ethics committee (id: 688). Before starting the data collection, the experimental setup was also preregistered on the Open Science Framework [165].

2.2.2 Recommender support system

For the study, we developed a system recommending the transcripts of similar previous chat conversations to a counselor based on the content of the counselor's current chat conversation. Figure 2.1 shows a chat window on the left and the support system interface on the right. The support system shows the top-10 most similar chat messages, which the counselor could click to read them in their entirety.

A corpus of chat conversations between help seekers and counselors was used to find similar previous chat situations. We used the corpus from 113 in the Netherlands. This corpus contained seven months of chats spanning from March 2018 to September 2018. The chat data were first filtered, removing all chats that had less than twenty interactions. In total, we used 17,773 chats. Furthermore, any special symbols in the messages were cleaned, and capital letters were replaced by lowercase letters.



Counsellor: What do you think of the idea to let a good friend know about your suicidal thoughts?

Counsellor: You think you could manage that?

Visitor: I don't want that

Counsellor: That's fine! What about just contacting your friends just to cope with your feelings?

Visitor: That would be alright

Counsellor: OK! How would you plan to do that?

Visitor: I could always just message my friends

Visitor: But I don't want to be a nuisance to them

Visitor: They are still young

Visitor: They should be able to enjoy their lives

Visitor: And not have to worry about me

Similar help-seekers

Counsellor: That could be the case, but I'm sure she would be willing to help you.
Visitor: I already said that I don't want to bother my parents with it
Visitor: sorry if I sound rude..
Counsellor: The more you try to not be a burder to your parents, the harder you are making it for yourself.
Visitor: My mom has enough problems as it is and if I tell them that I'm tired of my life and that I started smoking then she'll get super mad and I really don't want to live anymore then
Read more...

Visitor: No
Visitor: They want to be left alone
Visitor: and I love them so much *cry*
Counsellor: Do they know how you feel about not being able to talk to them about it
Visitor: no they think I'm overreacting
Read more...

Counsellor: Can you explain? How would you lose them?
Visitor: I'm afraid that if I tell them how I really feel, that I'll scare them away and that they don't want to deal with me.
Visitor: I don't want to put that burden on them
Visitor: I would rather talk about fun things
Counsellor: You want to be a good friend to them.
Read more...

Figure 2.1: Interface support system (right) and simulated chat (left). Content translated from Dutch.

Because the chats each contained multiple problems, we used a sliding window algorithm to scan for relevant chat segments instead of comparing complete chats. This algorithm created sets of chat messages, starting with the first five messages. The next set removed the first message in the window and added the sixth message; this process was repeated to create every possible set of five subsequent messages in a chat. The sliding window algorithm was then used to create the chat segments for the entire corpus.

We used an embedding algorithm to compute the similarity. For each chat segment, an embedding was created using smooth inverse frequency [72], which takes a weighted average of the word embeddings for each word in the text of the window corresponding to the inverse of the frequency of the word in the corpus. This resulted in less meaningful words receiving a lower weight. To create word embeddings, Mikolov et al. [42] developed an algorithm dubbed Word2Vec, improving previous methods [45]. The word embeddings we used were obtained from the COOSTO Word2Vec model [90], a model developed using Dutch social media and blog posts. A window of five messages resulted in 1,286,659 embeddings, which were stored alongside the corresponding chat and window positions.

When a counselor in an ongoing conversation requested similar chat conversations, a single smooth inverse frequency embedding was created using the same steps as with the corpus, except only the last five messages of the ongoing conversation were used. This embedding was then compared with the corpus embeddings through a cosine similarity. Ten windows with the highest similarity were recommended to the counselor.

2.2.3 Difficult chats

We used six chats for the experiment to cover several difficult situations: a situation where a help seeker 1) was in a dangerous location and had withheld this from the counselor; 2) did not want to inform anybody in their environment of their suicidality because they felt like it would put a burden on others; 3) was afraid of people in their environment not understanding their problems; 4) tried to look for help but was not believed; 5) was excessively rude; and 6) had to contact a psychologist.

2.2.4 Participants

Counselor and expert recruitment, as well as conducting the experiment, happened at 113 Suicide Prevention. In total, 24 counselors participated. On average, the participants’ age was 27 years old, and 79% were female. Only counselors who were interns, volunteers, or trainees were eligible to participate. Each counselor met all the components and conditions of the evaluation.

2.2.5 Measures

The perceived utility was assessed with the following question: "How, in your opinion, did the extra information help you with coming up with your response?". The counselors graded each support type on a fixed interval scale from ’3’ to ’3’ where ’3’ indicated the extra information was hindering, ’0’ indicated the information was neutral, and ’3’ indicated the information was useful.

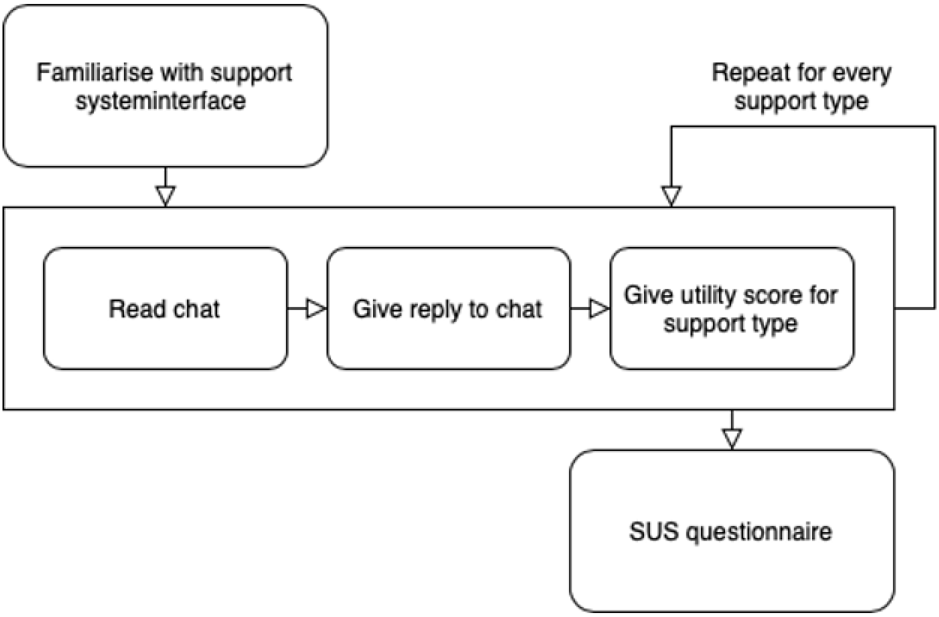

Familiarise with support systeminterface
Repeat for every support type
Read chat
Give reply to chat
Give utility score for support type
SUS questionnaire

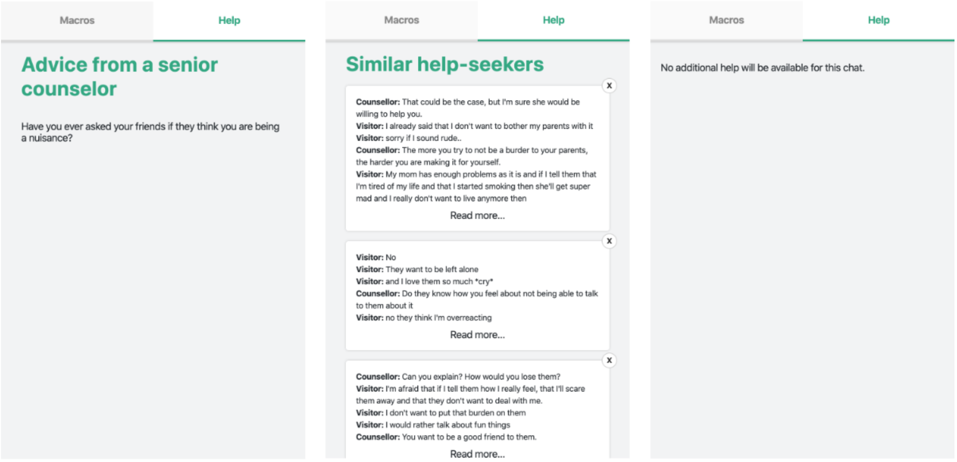
Figure 2.2: Procedure diagram of the first part of the experiment. SUS: System Usability Scale.

To measure usability, the counselors were asked to fill out the System Usability Scale questionnaire [6]; this is a validated ten-item questionnaire with a five-point scale ranging from "Strongly disagree" to "Strongly agree."

To measure the validity of the algorithm, the counselors used a seven-point fixed interval scale to indicate how much they agreed with the following statement: "The problem in the matched chat is the same as the problem in the ongoing chat." A score of ’1’ meant the counselor did not agree, whereas a score of ’7’ meant they did agree.

2.2.6 Procedure

The counselors used a test environment with simulated chats. The experiment consisted of two parts. Figure 2.2 shows a diagram of the procedure for the first part. Before the experiment, the counselor had five minutes to explore and familiarize themself with the support system.



Advice from a senior counselor

Have you ever asked your friends if they think you are being a nuisance?

Similar help-seekers

Counsellor: That could be the case, but I'm sure she would be willing to help you.
Visitor: I already said that I don't want to bother my parents with it
Visitor: sorry if I sound rude..
Counsellor: The more you try to not be a burder to your parents, the harder you are making it for yourself.
Visitor: My mom has enough problems as it is and if I tell them that I'm tired of my life and that I started smoking then she'll get super mad and I really don't want to live anymore then

Visitor: No
Visitor: They want to be left alone
Visitor: and I love them so much *cry*
Counsellor: Do they know how you feel about not being able to talk to them about it
Visitor: no they think I'm overreacting

Counsellor: Can you explain? How would you lose them?
Visitor: I'm afraid that if I tell them how I really feel, that I'll scare them away and that they don't want to deal with me.
Visitor: I don't want to put that burden on them
Visitor: I would rather talk about fun things
Counsellor: You want to be a good friend to them.

No additional help will be available for this chat.

Figure 2.3: Conditions of the experiment: senior counselor written advice (left); support system (center); no additional help (right). Content translated from Dutch.

Part 1 consisted of a simulated environment where the counselor read and reacted to three simulated chats, one after the other. The support information was contained in an extra tab called "Help." Figure 2.3 shows each support type. Each counselor had the same amount of time to read the chat. To simulate a real situation, each counselor had a two-minute window to reply. The counselor could not access the support tab before the two-minute timer started. Directly after the counselor submitted their reply to a chat, they were asked to rate the utility of the support type. These steps were repeated for each condition. Therefore, the participants reacted to three chats in total. The chats, support types, and combinations were counterbalanced for the 24 participants. This part ended with the System Usability Scale questionnaire being used to measure the usability of the support system.

Part 2 recorded the measurements for evaluating the validity of the algorithm. Figure 2.4 shows a diagram of the procedure. The left side of the screen contained the transcript of an ongoing chat. The right side of the screen showed ten chat segments. Half of these

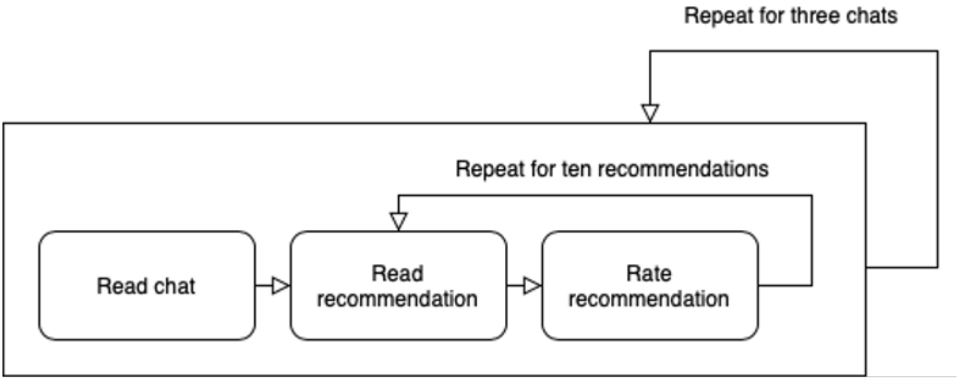


Figure 2.4: Procedure diagram of the second part of the experiment.

segments were randomly selected, and the other half was matched to the ongoing chat using the embedding algorithm. Below each of the segments was a fixed interval scale from one to seven where the counselor rated the degree to which that chat segment related to the ongoing chat. To enhance generalization, the participants did this for the transcripts of three different ongoing chats. Therefore, in total, a participant rated 30 segments.

2.2.7 Data preparation

Eight experts labeled the reply of the counselor with the type of help (condition) that the expert assumed that the counselor had received. To prevent expert bias, each expert judged all the counselor responses. Furthermore, a reliability analysis for the items of the System Usability Scale questionnaire showed an acceptable level of consistency, with a Cronbach alpha of .89. Therefore, the System Usability Scale items were compiled into a single score.

2.2.8 Analysis

The noticeable difference in counselor outputs was analyzed using generalized mixed-effects analyses [51] to predict the outcome variable support type based on the label the expert assigned to the counselor reply. The analyses were done by comparing two support type conditions at a time, thereby excluding the data from one of the three support type conditions. The models fitted on the remaining two conditions hence assumed a binomial distribution. Each model was compared with a null model that did not include an expert label as a fixed effect. Because the test was conducted three times, a Bonferroni correction [20] was used to set the significance threshold at .016. In addition, crossed random effects were used with random intercepts for the counselor and expert. Furthermore, for each support type, the utility ratings were analyzed using a one-sample t test to examine whether the rating deviated from the neutral zero score on the scale.

To examine the validity of the algorithm, a linear mixed-effects analysis was performed on the counselor's rating of the similarity between the chat segment and the ongoing chat. As a two-level fixed effect the analysis included the recommendation method, that is, randomly selected versus selected by embedding algorithm. Furthermore, the ongoing chat was added as a three-level fixed variable because the quality of the suggestions was assumed to depend on the specific chat. As a random effect, the intercepts for counselors were used.

Anonymized data and R scripts are available online [145].

2.3 Results

2.3.1 Noticeable difference in counselor output

Table 2.1 shows the effect of support type on the outcome measure of the expert label. The first row shows that the expert label significantly predicts the support type, when the data of no support condition was left out. In other words, the experts could tell the difference between replies given with the support system and replies with help from a senior counselor. Table 2.2 shows that the odds that the counselor received the senior counselor condition were 0.47 times less likely when the expert labeled the counselor's response as having received help from the support system. This effect is further illustrated when looking at the confusion matrix of these conditions, as shown in Table 2.3. However, no significant difference

Table 2.1: Results of the comparison between null model and full models that included the expert label as a fixed effect to predict support type counselors had received when writing their reply ($N = 356$)**.**

Outcomes data included in analysis	χ^2(degrees of freedom)	p value
Support system and senior counselor written advice	11.31(2)	.004
No support and support system	1.44(2)	.49
No support and senior counselor written advice	4.78(2)	.09

Table 2.2: Fixed effect of the expert label for the model of support system and senior counselor written advice.

	Estimate	OR	Std. error	z Value	p value
Intercept	0.25	1.29	0.16	1.54	.12
Support system	-0.755	0.47	0.25	-3.04	.002
Senior counselor written advice	0.014	1.01	0.28	0.049	.96

was found between the no support condition and any of the other conditions.

2.3.2 Utility

The results of the utility ratings are shown in Table 2.4. The mean score of the support system was -0.21 (SD 2.26) and did not significantly deviate from 0, indicating that there was neither a hindering nor a helping effect experienced by the counselors. However, the mean utility score of the written advice from a senior counselor was 1.46 (SD 1.91) and significantly deviated from 0. This suggests that the written advice was perceived as helpful. It is noteworthy that the support system had a high variance, suggesting that the counselor's opinion on the utility was divided.

2.3.3 Usability

The mean score of the support system for the SUS questionnaire was 71, with a 95% confidence interval of 63 to 78. According to Bangor et al. [26], this score can be classified as "good" based on an adjective rating scale.

2.3.4 Validity of the algorithm

How the chat segments were selected (randomly vs by the embedding algorithm) significantly predicted the rating counselors

Table 2.3: Confusion matrix for expert labeling of counselor responses.

		Condition		
		No support	Support System	Senior counselor written advice
Expert label	No support	**74**	66	85
	Support System	65	**76**	46
	Senior counselor written advice	39	36	**47**

Table 2.4: One-sample T-test for counselor utility ratings per support types (n=24).

			95% CI				
Support type	**Mean**	**SD**	**Lower**	**Upper**	**t**	**df**	***p* value**
Support system	-0.21	2.26	-0.84	0.43	-0.68	23	.50
Counselor written advice	1.46	1.91	0.87	2.04	5.17	23	<0.001
No support	-0.21	0.95	-0.62	0.20	-1.04	23	.31

gave on the chat segment's similarity to the ongoing chat, $\beta = 0.30, t(7.66), p < .001$. This means that counselors could tell the difference between the random chats and those generated by the support system. The suggestions from the algorithm increased the similarity rating given by counselors from an average of 2.35 to an average of 3.42 (difference of 1.07).

2.4 Discussion

In the current study, we evaluated a prototype support system to assist chat counselors in suicide prevention helplines by providing inspiration from previous chats. The results show that counselors gave different answers depending on whether they received help from the support system or from a senior colleague. Upon inspection, the replies given by the counselors who received written advice from a senior colleague were, for the most part, copied directly and with little to no alterations made. Replies from counselors using the support system were more varied. This could be a possible explanation for the noticeable difference. However, we could not find a significant result for the no-help condition, which also had varied replies. Additionally, we observed that written advice from a senior counselor was given a significantly higher utility score than the other conditions; this suggests that the counselors value short actionable information that is highly accurate to the situation and that is given by someone with expertise. Gunaratne et al. [54] have observed similar findings in their study on the effects of expert advice and social comparison on decision making for retirement savings; they showed that expert advice helped people make better decisions, whereas social comparison was seen as a useful mechanism to keep people from deviating too far from the mean and, hence, make safe decisions. However, both of these conditions outperformed a control condition where no additional information was provided.

The main contribution of our study is the idea of retrieving inspiration from a conversation corpus. Other support systems for chats [12, 21, 93] have used topics to assist the conversation. Compared with these methods, our approach for combating writer's block in a counseling conversation is novel. Furthermore, an experimental design was used to compare this supportive technology with advice from a senior colleague, showing how the two differ.

Some limitations should be considered regarding the findings and their implications. We used chat transcripts of conversations with situations that previous counselors found difficult to evaluate. However, this might not cause writer's block for every participant because not every counselor will have problems with the same situations. For writer's block to occur naturally, the system would have to be tested in live chats. This was, however, not possible because of the ethical constraints of deploying an unevaluated system in a possible crisis situation. Furthermore, the specification, development, and evaluation were done in the context of counselors working at 113 in the Netherlands, with a limited number of counselors. The support system should also be tested in different helplines and with a larger sample size.

We have outlined two major directions for future work. First, the recommendation mechanism could be improved in different ways. This study, as well as other related works such as recommenders for creativity [34] and scientific writing [61], relies on topic modeling and bag-of-word models to find recommendations. Encoding text using attention-based models [180], such as BERT [98], have been shown to perform well on various natural language processing tasks, including semantic sentence similarity for conversation data [93]. These methods could be applied to improve the recommendations to find more relevant and similar examples, which we expect will increase the perceived utility. Additionally, curating the corpus can help denoise the dataset and improve the recommendations. This can also give counselors the knowledge that the information comes from a subset of quality chats, thereby acting on the persuasive principle of authority as outlined by Cialdini [21]. Lastly, there is also an opportunity to apply active learning methods by adding positive labels to the recommendations that the counselors interacted with or explicitly marked as useful [12].

Second, the findings show that the embedding algorithm found similar chats and that written advice from senior counselors had high utility. Compared to the Gunaratne et al. study [54], the main difference to the setup of our study is that the social comparison condition provided information as an average; this indicates that refining the output of the support system recommendations to be more instructional could be a possible direction for improving the system. To combine both the extensive coverage of a chat corpus and the high utility of curated written advice, clustering could be used, that is, grouping similar chats together based on a similarity metric and curating the labels based on these clusters. Derrar [97] uses clustering to automate the annotation of customer service chat messages. A similar approach could be used to annotate the chat corpus to create a taxonomy of situations and advice, which then could emulate receiving written advice from a senior colleague. In other words, working together with experts, a set of advice could be formulated in advance for each specific situation. Next, a data-driven algorithm could be trained to classify chats according to categories of the taxonomy, consequently providing counselors with expert advice associated with the category and making the expert advice situation relevant. This approach would be most suitable for assisting counselors with frequently occurring tasks, as these would be the most likely cases to be included in the taxonomy. The focus of the support system might therefore shift from an inspiration source to a system that could reduce workload. Alternatively, the field of conversational information retrieval has explored multiple methods that could be applied to the task presented in this paper. For example, Qiu et al. [80] combined both information retrieval methods and generation based models to create a chat bot trained using existing customer service chat logs. These techniques could potentially also be used to allow the system to generate proposal responses that counselors could consider using in their chats with help seekers.

2.5 Conclusion

In conclusion, the current study shows a possible method to provide inspiration during chat counseling in a helpline for suicide prevention and how this supportive technology compares with human assistance. A support system may be a relief for counselors as they handle many cognitively difficult situations. In addition, supportive technology seems useful for helplines to better deal with busy periods, to provide a safety net for junior counselors, and to help sustain counselors.

Chapter 3

Evaluation of a deep learning-based helpline support tool

Salmi, S., Mérelle, S. Y. M., van Eijk, N., Gilissen, R., van der Mei, R. D. & Bhulai, S.

Submitted for publication.

Abstract

Background

This study aimed to evaluate the effectiveness of a deep learning-based recommender system in enhancing counselors' self-efficacy during suicide prevention helpline conversations. A secondary goal was to assess the quality of the suggestions provided by the tool and the resulting conversations.

Methods

A randomized controlled trial was conducted with counselors assigned to either a control group or an intervention group. The intervention group used an AI-assisted tool during shifts, which generated suggestions from successful past sessions using a transformer-based architecture and Sentence-BERT embeddings. The control group received usual care. Both groups completed a self-efficacy questionnaire after each shift, for up to ten shifts.

Results

In total, 48 counselors participated: 27 in the experimental condition and 21 in the control condition, rating a total of 188 shifts. No significant difference in self-efficacy was found between the conditions ($p = 0.36$). Counselors using the AI tool had slightly lower response times and used the tool more frequently during long conversations. The tool was often used after a response had already been given and was only used correctly in 64 conversations. When used correctly the tool provided usable information in 83% of cases.

Conclusion

The deep learning-based recommender system provided real-time, usable suggestions during counseling sessions. While there was no significant impact on self-efficacy, the tool's frequent use in longer conversations suggests counselors found it helpful in complex situations. The study shows the feasibility of integrating AI-assisted tools in suicide prevention to improve counselor support and helpline service quality when scaled up.

3.1 Introduction

In Chapter 1, taxation of counseling was introducted as one of the current challenges helplines deal with. Several studies have explored tools to assist counselors in their online work. Dinakar et al. [52] provided live topics through topic modeling. In the previous chapter a support tool was proposed that provided suggestions for counselors. This tool aligned with a proposed support tool framework by Madeira et al. [122], who identified that a potential support tool would use NLP to classify chat messages and then provide suggestions. Another approach was used by Demasi et al. [96], who proposed a retrieval-based chatbot that acts as a training partner for counselors.

This chapter builds upon the work in Chapter 2 by addressing its main limitations. The previous study used a retrieval algorithm based on word embeddings and was evaluated on mock conversations. In this work, we used deep learning to train a support tool on a large amount of helpline conversation data. By doing so, we aimed to create a tool that offers better embeddings and, thus, better suggestions while being implemented in real-time conversations. This was expected to boost the counselors' confidence and enhance their ability to support people in distress effectively.

Due to its complex nature, text retrieval in a suicide prevention setting is an ideal use case for transformer-based approaches. Methods like cross-encoders allow for effective sentence similarity, but are slow in real-time use. If text is converted to a single embedding, it can be compared efficiently to a large database of embeddings through basic distance metrics, like the cosine distance. This approach has been shown to work well when comparing text embeddings [65] [108]. A popular method used in retrieval-based dialog systems is to combine these methods [154]. A top *n* number of documents are preselected as candidates using a less powerful but computationally less intensive method, like cosine similarity. Then afterward, the list of candidates is reranked using a more powerful but computationally more intensive method like cross-encoders [169]. For our approach however, we opted only to apply sentence embeddings and have the counselor be the final selector of the top *n* list instead. This has the benefit of having a faster and simpler system, while giving the counselor more autonomy. This also has the benefit that the counselor can combine inspiration from multiple suggestions.

It is important, however, to have good sentence embeddings. A simple method to get an embedding for a sentence, or small piece of text, is to average word embeddings within a text. This method can, however, only capture a limited selection of helpline cases. To be able to include as many helpline situations as possible, we therefore looked to more tailored methods that were available.

One of the prominent approaches for sentence embedding is Sentence-BERT, introduced by Reimers et al. [108]. This method builds upon RoBERTa [105] by extending it to generate sentence embeddings. It employs a Siamese RoBERTa network and is trained using triplet loss. The resulting embeddings can be efficiently compared using cosine similarity.

In this work, we trained a sentence embedder in the context of a chat dataset. We generated training samples in an unsupervised manner and fine-tuned a pre-trained transformer network using a variation on the sentence embedder from [108]. Furthermore, we evaluated this tool, with a primary goal of observing the impact on counselors' self-efficacy after a shift. To investigate this impact, counselors used the support tool in real-time for 188 chat conversations. We assessed counselors' self-efficacy with a randomized trial. A secondary goal was to evaluate the suggestions provided by the tool to the resulting conversations that they were use in. To our knowledge, this study is one of the first applications of a support tool to provide conversation suggestions in a real-time setting.

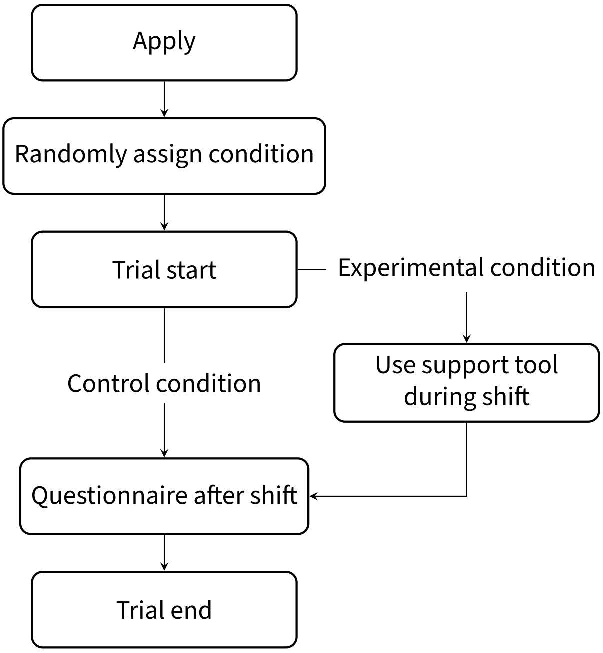

Apply
Randomly assign condition
Trial start
Experimental condition
Use support tool during shift
Control condition
Questionnaire after shift
Trial end

Figure 3.1: Experiment design

3.2 Methods

3.2.1 Study design

This study was designed as a randomized controlled trial. Counselors were randomly assigned to a control or intervention group. The intervention group was granted access to the support tool during their shifts, while the control group did not have access to it (care as usual). At the end of ten shifts, both groups were asked to fill in a four-item questionnaire. Figure 3.1 describes the workflow of the study design. Furthermore, the usage of the support tool during chat conversations was monitored.

3.2.2 Participants

Participants were recruited from 113 Suicide Prevention in the Netherlands. Initially, counselors were recruited through a questionnaire, where they could provide a code of their own choosing. To remain anonymous, this personal code was used to access the support tool and fill in the efficacy questionnaires. After the recruitment period of one month the participating counselors were randomized, accounting for gender and age. Due to a low response rate, we deviated from the initial recruitment method, and used another recruitment method for a period of two weeks. A randomized list of unique codes was generated to ensure participant anonymity, with a predetermined allocation to each condition. Counselors were asked to participate in this research during daily meetings at the start of their shifts. A helpline manager would sequentially assign participating counselors based on this randomized list. Counselors used the provided code from the list to access the support tool and fill in the questionnaires. The manager retained the code list with corresponding participant names, so they were aware of which counselors were using the tool, and could provide reminders and instructions for the questionnaire.

3.2.3 Measurements

Primary outcome measure

Participants were asked the question to rate on a Likert scale ranging from ’1’ to ’10’ their response to the statement "I was able to handle difficult situations during this shift".

Covariates

Besides the primary outcome measure, we included two covariates.

- Participants were asked if they sought assistance from the floor manager during their shift. They could indicate if they requested no assistance, some assistance, or a lot of assistance during their shift.
- Participants were asked to indicate the type of shift they had. They could indicate if they had a morning, day, or evening shift. Night shifts were not included as the helpline protocol is different from the other shifts.

Secondary outcome measure: support tool usage

Participants were asked whether they utilized the support tool during their shift. They could indicate they did not use, sometimes used, or frequently used the support tool during their shift.

In addition to the questionnaires, we assessed the counselors' usage of the tool by examining the frequency and duration of their interactions, using the date and time of the messages in the corresponding conversations. Moreover, we labeled the interaction for each instance when a counselor accessed the support tool. This labeling was carried out separately by the researchers SS and NvE and encompassed two primary aspects: first, we determined whether the support tool provided applicable information, and second, whether the counselor actually incorporated information from the support tool. If the researchers disagreed with the label, the situation was examined and discussed by both authors until a consensus was achieved.

Additionally, we observed instances where counselors accessed the support tool at potentially inappropriate times, seemingly for testing purposes. We identified four such cases: when the tool was queried after the counselor had already responded, when the help seeker was in the midst of composing multiple messages and had not yet finished, and when the counselor queried at the very start or the very end of a conversation. Table 3.1 shows the labels and their definitions.

3.2.4 Statistical analysis

A mixed-effects analysis was used to predict the primary outcome of the statement "I was able to handle difficult situations during this shift" based on the group to which the counselor belonged (experimental or control condition). In this context, the outcome of this question served as the dependent variable, and the group represented the independent variable. The type of shift (morning, afternoon, evening) and if they received assistance from a floormanager were included as covariates.

Table 3.1: Label definitions

Label	Options	Definition
Usage	Adopted	One of the five suggestions was adopted or copied.
	Elements	Elements of one or more of the five suggestions were used.
	Disregarded	None of the five suggestions were used.
Usability	True/False	There was useable information among the five suggestions.
Incorrect usage	Early	The help seeker wrote a message after the tool was queried but before the counselor could answer.
	Late	The counselor wrote a reaction to the help seeker and queried the tool afterward.
	Double	The counselor queried twice without change in the conversation.
Incorrect situation	Begin	The counselor queried the tool at the beginning of the conversation or right after it was transferred to them.
	End	The counselor queried the tool at the very end of the conversation.

3.2.5 Support tool

This section provides an overview of the support tool pipeline. We will describe the architecture of the network to obtain the sentence embeddings and elaborate on the training method. Subsequently, we discuss how inference is done when a user requests a recommendation. We show the user interface and demonstrate how a counselor can interact with the tool. Finally, we describe the dataset that was used to train the support tool.

Paired triplet loss network

To create a text embedder for the helpline conversations, a deep learning method is used with a modification on the triplet loss strategy used by [108]. The triplet loss function is defined as:

$$triplet(A,P,N) = \max(d(A,P) - d(A,N) + margin, 0)$$

where d is a distance function, A is an anchor sample, P is a positive sample (i.e., a sample of the same class as the anchor) and N is a negative sample (i.e., a sample of a different class than the anchor). Through this loss, the network is encouraged to embed samples such that the distance between the anchor and the positive sample is smaller than the distance between the anchor and the negative sample by at least the defined margin. The distance function d is learned through a transformer network.

However, the dataset only contains chat conversations with no indication of positive and negative classes. To obtain these positive and negative samples to fine-tune the network, we generated anchor and positive sample pairs from the chat conversations and relied on randomization to retrieve negative samples. We randomly use a positive sample from a different anchor positive pair.

To create these samples, a conversation was split after each help-seeker message. This resulted in two texts: (1) the conversation preceding the split and (2) the conversation following the split, which we used as the anchor and the positive samples respectively. This split was done after each help-seeker's message in a conversation. For example, given a sequence $\{c_1, h_1, c_2, h_2, c_3, h_3\}$ of counselor messages c and help seeker messages h, we can create the pairs: $S_1^b = \{c_1, h_1\}, S_1^a = \{c_2, h_2, c_3, h_3\}$ and $S_2^b = \{c_1, h_1, c_2, h_2\}, S_2^a = \{c_3, h_3\}$.

Formally, let $D = \{m_1, m_2, \ldots, m_n\}$ be a document consisting of n messages. Define the sets $S_i^b = \{m_1, \ldots, m_i\}$ and $S_i^a = \{m_{i+1}, \ldots, m_n\}$ for $1 \leq i < n$. Then the set of all possible pairs (S_i^b, S_i^a) is given by:

$$\{(S_i^b, S_i^a) | 1 \leq i \leq n-1, m_i \in H\}$$

where H is the set of messages from help seekers.

To provide a way for the network to identify which message originated from which conversation actor, we added two additional special tokens besides the class and separator tokens native to most BERT-based networks. Each message in a set was prepended with either a token for a help seeker or a token for a counselor, denoted "<HLP>" and "<CNS>" respectively.

The messages from each S_{ij}^b and S_{ij}^a were concatenated to create two texts, T_{ij}^b and T_{ij}^a respectively. Due to the size limitation of transformers, when a text exceeds the token limit of the RoBERTa network, truncation needs to be applied. To preserve as much relevant text as possible, each T_{ij}^b was truncated from the left side, while each T^a_{ij} was truncated from the right side.

We used the average of the final hidden state of a pre-trained RoBERTa network to embed the texts. We found that separate models instead of a Siamese approach provided slightly better representations when comparing embeddings produced by the same network. In the envisioned support system, only texts of T^b_{ij} would be compared. We, therefore, opted for separate models, B^b and B^a. This resulted in embeddings $E^b_{ij} = B^b(T^b_{ij})$ and $E^a_{ij} = B^a(T^a_{ij})$.

We apply the triplet loss function on the E^b_{ij} and E^a_{ij} pairs. To ensure that each E^a_{ij} is used for a positive as well as a negative example, we combine two pairs such that

$$loss(E^b_{ij}, E^a_{ij}, E^b_{lk}, E^a_{lk}) = (triplet(E^b_{ij}, E^a_{ij}, E^a_{lk}) + triplet(E^b_{lk}, E^a_{lk}, E^a_{ij}))/2$$

Training

The RoBERTa model we used was pre-trained on a Dutch language corpus [118]. We used the L^2-norm for the distance function and 1 for the margin. The models were trained for two epochs with a learning rate of 1×10^{-5}.

Inference and retrieval

After training, the B^b network was used to embed all T^b_{ij} texts. These embeddings were stored alongside their corresponding positions. During a conversation, a counselor could query these stored positions through a support tool. The tool scraped the current conversation and submitted the last several interactions. These messages were also prepended with special tokens and concatenated into T_h, identically to the training approach. From this text, we obtained embedding $E_h = B^b(T_h)$. E_h was compared to each stored embedding E^b_{ij} using the Cosine Similarity. Using the top five embeddings with the highest similarity score, we obtained the corresponding document D_j and position i which were used to show the relevant messages to the counselor.

Interface

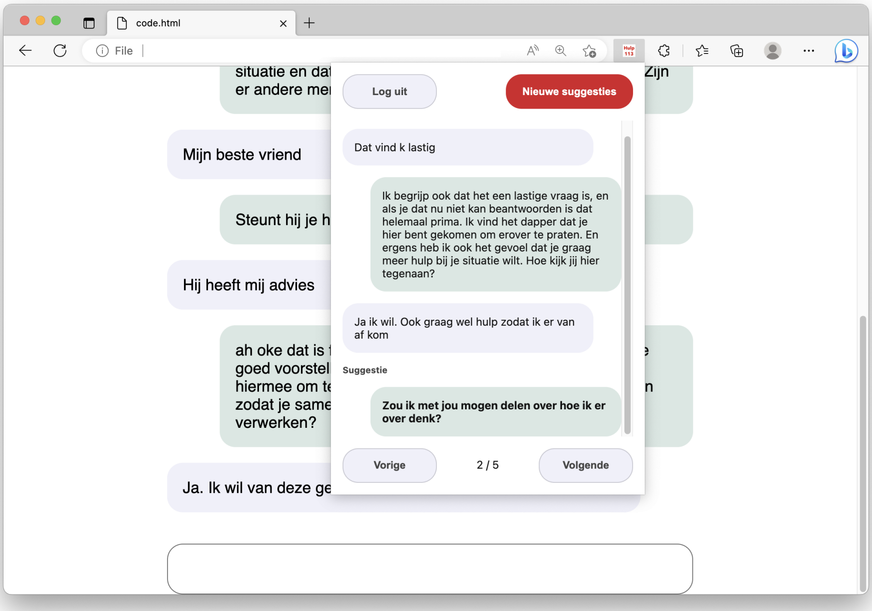
The suicide prevention helpline utilized a browser-based chat program as a means of communication with help seekers. The support tool was implemented as a browser extension that functioned by analyzing the conversation when counselors requested support. The chat information was transmitted to a server, where embedding and retrieval processes took place. The resulting information was then sent back to the counselor and displayed within the tool.

Figure 3.2 showcases the user interface of the tool. The tool presents the counselor with five alternatives, corresponding to the top five most similar embeddings determined by Cosine Similarity scores. The three preceding messages leading to the suggestion were also displayed to allow for comparison with the counselor's conversation. The suggestion itself was clearly labeled at the bottom. The counselor could cycle through the five suggestions using the previous and next buttons. If the conversation progressed and the counselor desired new suggestions, they could request them by clicking a dedicated button. It is worth noting that the tool intentionally disabled the copy-and-paste functionality to encourage counselors to rephrase the suggestion to align with their individual conversation style.

Support-tool database

The dataset utilized for the database of the support tool consisted of chat interactions from previous chat conversations from the 113 Suicide Prevention, the national Dutch suicide prevention helpline. These chat interactions included both pre-chat and post-chat information filled in by the help seekers. Notably, only chats that received a positive rating from the help seekers were included in the dataset, ensuring a focus on successful counseling sessions.

Chat conversations were anonymized and help seeker and counselor names were given generic tokens. This was done by replacing usernames in the dataset. Furthermore, the deduce library for anonymization of clinical texts was used to further anonymize other



situatie en dat
er andere men

Zijn

Mijn beste vriend

Steunt hij je h

Hij heeft mij advies

ah oke dat is
goed voorstel
hiermee om te
zodat je same
verwerken?

Ja. Ik wil van deze ge

Dat vind k lastig

Ik begrijp ook dat het een lastige vraag is, en als je dat nu niet kan beantwoorden is dat helemaal prima. Ik vind het dapper dat je hier bent gekomen om erover te praten. En ergens heb ik ook het gevoel dat je graag meer hulp bij je situatie wilt. Hoe kijk jij hier tegenaan?

Ja ik wil. Ook graag wel hulp zodat ik er van af kom

Zou ik met jou mogen delen over hoe ik er over denk?

Figure 3.2: Interface of the support tool (in Dutch)

components, such as locations and names of third parties.

The final dataset comprised a total of 187,000 messages exchanged during 3,816 chat sessions. From these messages, 91,252 unique embeddings were generated, representing potential suggestions that can be provided by the support tool.

Ethics approval

The study protocol is performed in accordance with the relevant guidelines. This study was reviewed and approved by the Medical Research Ethics Committee of Amsterdam Universitair Medisch Centrum (registration number: 2022.0855).

3.3 Results

3.3.1 Study sample

For the initial recruitment period, 40 counselors signed up to participate. After the start of the trial only fifteen counselors filled out the questionnaires. The second recruitment included an additional 33 participants who filled out the questionnaires. In total we had 48 participating counselors in the experiment, 27 in the experimental condition and 21 in the control condition. We aimed for 37 counselors per condition for an appropriate power; however, unfortunately, this number of participants was not reached. The flow of the participants through the trial can be seen in Figure 3.3. This resulted in a total of 167 completed questionnaires, 93 in the control condition and 74 in the experimental condition.

3.3.2 Primary outcome: Counselor self-efficacy

We did not find a significant difference in the self-reported score for the counselor's ability to handle difficult situations ($p = 0.36$). The support tool condition scored an average of $7.5(SD = 2.2)$, and the control condition scored an average of $7.9(SD = 1.9)$.

3.3.3 Secondary outcomes

Response time and conversation duration

The response time of counselors using the support system is $43.75(SD = 11.91)$ seconds, which is slightly lower than $47.34(SD = 27.25)$ for regular conversations, but not significantly different ($p = 0.32$). The support tool is used in longer conversations ($p < 0.001$), lasting on average $66.22(SD = 19.13)$ minutes, compared to regular conversations, which last on average $52.19(SD = 26.42)$ minutes.

Support tool performance

Table 3.2 illustrates the coding of instances in which the support tool was accessed. The results showed that the support tool was used in 188 situations, in which eight were corrupted data points leaving 180 total labeled instances. Excluding irrelevant situations involving tool queries at the beginning (greeting) or end (goodbye), likely representing moments where the counselor was testing the tool, 152 queries remained. Among these, 103 (67%) received useful responses from the support tool.

However, this includes situations where the counselor queried the tool after providing their own response or while the help seeker

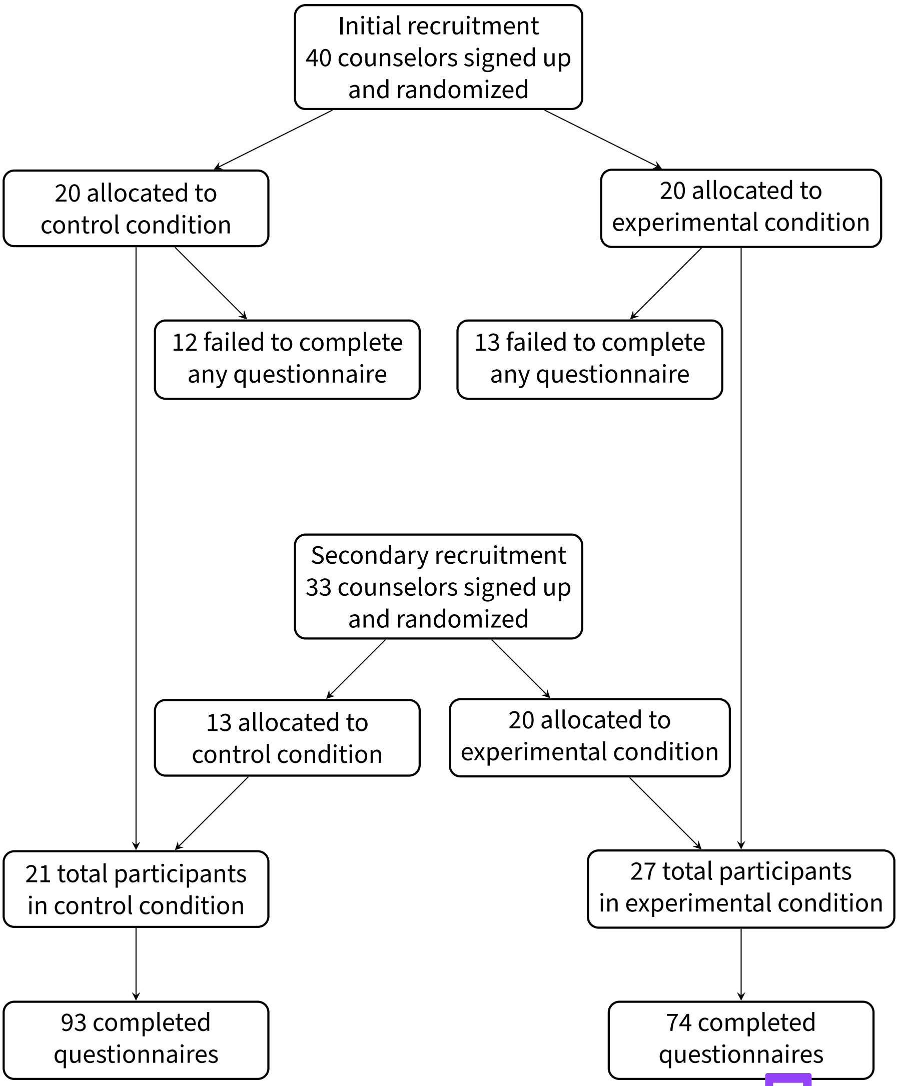

Initial recruitment
40 counselors signed up and randomized
20 allocated to control condition
20 allocated to experimental condition
12 failed to complete any questionnaire
13 failed to complete any questionnaire
Secondary recruitment
33 counselors signed up and randomized
13 allocated to control condition
20 allocated to experimental condition
21 total participants in control condition
27 total participants in experimental condition
93 completed questionnaires
74 completed questionnaires

Figure 3.3: Flow of participants through the trial

Table 3.2: Labeled support tool suggestions. Shows the number of useful and not useful suggestions, as well as the suggestions that were copied, taken elements from or where the counselor used something different.

	All situations	Relevant situations	Relevant situations and correct usage
Useful	113	103	53
Not useful	67	49	11
Copied	28	25	15
Elements	36	34	21
Different	110	93	28
Total	180	152	64

was still typing. When these were filtered as well, it resulted in 64 remaining queries, of which 53 (83%) were deemed useful. There was a noticeable gap in chat situations where the tool was queried at the appropriate time. In many cases, the support tool was queried after the counselor had already responded, suggesting a potential lack of initial trust in the tool. Out of the 53 instances where the tool was considered useful, 15 (28%) involved counselors copying the suggestion, and 21 (40%) incorporating elements from it and in the remaining 17 (32%) instances the counselor used a different response.

Recurring situations

The tool was frequently used in situations where the conversation was passed from triage to counselor. This could be because the counselor was testing the tool. However, the first step in any conversation is for the counselor to read the triage conversation, and it could therefore also indicate that there might be a demand for a summarization feature.

After further inspection, the cases in which the support tool may find the best recommendations can be described as messages that are relatively short and concise, such as questions or remarks. Furthermore, when the scenario frequently occurs in the corpus it would be easier to find better recommendations. Similarly, when the messages are longer, or describe specific and personal situations, then it would be more difficult to find good recommendations.

In the case where the counselors adopted the recommendation in their response, the types of responses were very broad, and covered many different situations. One common situation that was queried frequently was a scenario where a help-seeker was reluctant to talk about their suicidality to people in their environment. Furthermore, we saw that some counselors would adopt recommendations from the tool more frequently than others, indicating there might be a personal component.

3.4 Discussion

3.4.1 Principal findings

In this study, we developed and evaluated a retrieval-based support tool designed for use during counseling chat conversations. Not many similar tools exist, and this is a good step towards bringing real-time artificial intelligence to helplines. Our analysis looked at self-efficacy, response time, conversation duration, and support tool performance. The findings provide valuable insights into the tool's impact on counselor performance and user engagement.

Whereas many attempts were made to use AI as a counselor [73] [104], they had difficulty performing at the standards of mental health professionals [113], which created ethical concerns [167]. Our study shows a possible application of assisting humans with AI.

The support tool did not find an overall increase in the counselors' confidence in handling difficult situations during their shifts. We did, however, see that, in the intended use case, it produced a high percentage of relevant content. Furthermore, the longer duration of discussions may indicate that the tool was used in deeper or more difficult conversations. We observed that there was no significant impact on the response time of counselors, indicating it did not make the counselor too slow to respond to the help-seeker. The observed gap in tool usage at the appropriate time is something that needs to be addressed. Possibly, this can be mitigated in the future by building trust by improving the suggestion retrieval capabilities, combined with training and familiarization of counselors with the tool. Encouragingly, when used correctly, the tool demonstrated a high percentage of usability, with counselors adapting two-thirds of these usable suggestions. The remaining third of instances that were not adopted could be explained by counselor style, or because it might not have fit the current conversation stage. These features are difficult to capture, however. We found short and concise questions to provide the most consistently good responses. This feature of the support system would be less suited for in-depth therapy sessions, but could be very useful for assisting new counselors in getting accustomed to these situations faster.

Our study aligns with the current advancements in data analytics, NLP, and deep learning, aiming to enhance online mental health helplines. Our findings lay the groundwork for AI assisting helpline professionals. Ethical concerns [99] about harmful instructions are addressed by using only human-generated content, within its context, and by always having a human counselor making the final decision on what to write. This work underscores the potential of AI-based systems to assist counselors in navigating challenging helpline interactions, contributing to the broader goal of reducing suicides and suicide attempts.

3.4.2 Limitations

It is essential to acknowledge the study's limitations, such as the relatively small sample size and potential biases in counselor participation. Future research could explore interventions to address the observed lack of initial trust in the tool, potentially through targeted training or interactive sessions. Further enhancements to the tool, including a summarization feature, could be considered based on identified patterns of usage. As observed in the results, when long descriptive scenarios are given by the help seeker, it is difficult to find relatable content at a glance. This is a limitation of the support tool, because these situations can also cause writer's block. We found that long text input finds messages of equal length, which takes more time for the counselor to read through, so it is not immediately apparent if the suggestion is relevant. And even if there are common elements between the longer texts, there is no guarantee that the responses reference these common elements.

3.4.3 Implications and future work

The observed characteristics of good performance in frequently occurring situations and short to medium-length responses lend itself as a useful means to assist new counselors. The tool could be introduced in the training, in order to increase the trust in the system. An interesting avenue for future work would be to compare this application of a recommender support tool to a generative support tool. Using the recently popularized Large Language Models (LLMs), a model could be trained on counseling data to generate recommendations instead of retrieving them from a database.

While counselors did indicate in previous research that they would have trouble trusting AI tools that generate information, it would be interesting to see if this sentiment remains with how much the capabilities of LLMs have improved. This approach would also address one of the limitations of finding relevant content through the long descriptive scenarios that help seekers give. However, this approach must be taken with care, as thee reliability of LLMs in the mental health field is still a topic of discussion [172]

In the short term, the support tool can already provide feedback during a decent percentage of the requests of difficult situations. With the findings of this study, helpline counselors can be better instructed in the optimal times to use the support tool. Furthermore, the barrier to access can be lowered, by removing some of the necessary components for the study, such as the login codes.

3.5 Conclusion

In summary, this research explored the impact of a retrieval-based support tool in counseling chat conversations. While it did not significantly affect counselors' self-efficacy, and the usage was relatively low, the tool was highly relevant in most situations where support was needed. On average the support tool was used in chats of longer duration than normal chats. This led us to believe the tool was possibly used as a support mechanism to alleviate mental fatigue, acting as a resource to assist counselors coping with the extended cognitive demands of lengthy conversations. Despite limitations, the study lays the groundwork for advancing real-time AI in helplines. Future research should explore interventions to address initial trust issues and consider enhancements like a summarization feature. Comparing the current approach to a generative tool using large language models presents an intriguing avenue for future investigation, contributing to the ongoing refinement of mental health support tools.

Part II

Topic modeling for helplines

Chapter 4

Evaluation of topic modeling methods for conversations

Salmi, S., van der Mei, R. D., Mérelle, S. Y. M. & Bhulai, S.

Telematics and Informatics Reports, **13** 100126 (Mar. 2024)

Abstract

Background
Traditional topic modeling methods like Latent Dirichlet Allocation struggle with short, context-dependent texts typical of mental health helplines. Newer methods, leveraging pre-trained embeddings and variational inference, offer better performance by incorporating sentence-level context.

Methods
This study evaluates LDA, GSDMM, BERTopic, TopClus, and CombinedTM on three datasets: a suicide prevention helpline corpus, the AnnoMI mental health interviews, and the Multi-Domain Wizard-of-Oz dataset. Document representations tested include full conversations, individual utterances, and segments of utterances. Sentence embeddings were generated using SBERT, with fine-tuning using a Dutch RoBERTa model and a Transformer-based Denoising AutoEncoder.

Results
BERTopic, utilizing sentence embeddings, outperformed traditional methods, showing superior topic coherence and word embedding similarity, particularly on utterance-level and segment-level data. Hierarchical clustering methods like HDBScan handled noisy data effectively. Fine-tuning pre-trained embeddings provided marginal improvements.

Conclusion
Incorporating sentence embeddings significantly improves topic modeling in conversation data. BERTopic is particularly effective for mental health conversations, offering valuable insights. Future research should include data from multiple helplines and explore hierarchical models for better generalizability.

4.1 Introduction

Topic modeling is a common method to extract latent semantic information from text in documents. It is often used to gain insight into a corpus of some kind, for example, what the current events are on social media. LDA is the most popular method for topic modeling [14]. Many alternatives have been proposed that aim to outperform LDA in general cases, or for special kinds of text content such as short text.

In this chapter, we highlight one of the special kinds of corpora, namely conversations. Specifically, we look at conversations that are characterized by two things: First, information that is frequently locally contextual, and second, utterances that do not conform to a recurring topic. Such a corpus conflicts with the assumption made by LDA and other BoW models, namely that the order of words does not matter. For example, online mental health services or eHealth could use topic modeling to gain insight into their conversation data. Using more recent methods, topic modeling has been applied in counseling conversations on suicide prevention helplines [166].

Especially in long conversations, word-document co-occurrence lacks the information to describe the topics accurately. The resulting topics become too general or not coherent. In structured documents, one could split the document at some point in the hierarchy to reduce the size, for example, per chapter of a book [15].

Similarly, a conversation can be viewed as a collection of utterances between participants, where an utterance is a statement or message from one of the participants. It is more likely that participants of the conversation cover a small number of topics at a time. By splitting a conversation into its utterances, one can reduce the number of topics per observation. In this way, the co-occurrence of words has more descriptive value. We hypothesise that for conversations, especially mental health conversations, where descriptive value of words is already low, splitting the conversation into utterances improves the topic model coherence. However, splitting the conversation increases the sparsity of the data. Due to the nature of statistical inference used by many topic modeling methods, this is a problem.

Several topic methods have tried to move away from the bag-of-words assumption and take sentence-level context into account [23, 32, 37]. Tian et al. [68] proposed the so-called Sentence Level Recurrent Topic Model to create distributed representations of sentences and topics using a Long-Term Short-Term neural network. Sentence embedding is a method to find a latent representation of a sentence in a continuous and lower-dimensional space. The most successful sentence embedding methods rely on the recent transformer model [180]. This model has shown improvement in many natural language processing tasks [98, 105, 135]. Sentence embedding can be applied to topic modeling to extract more information than simple word co-occurrence.

A downside of using sentence embedding for topic modeling is that the embedding can only be as long as the maximum input length of the model. This means that for long documents, words after the maximum length are truncated. However, conversation utterances are small enough that this maximum length is unlikely to be reached. For this reason, we believe that sentence embedding methods can be of benefit for utterance level conversation data.

A middle ground between utterances and a full conversation is to segment a conversation into groups of utterances. However, this creates an additional problem, namely: how should the conversation be segmented? A frequently used method for the purpose of text segmentation modifies the TextTiling algorithm [7]. This method employs embedding models based on BERT to embed or classify sentences to provide a lexical score for each sentence [147, 151, 173]. Lexical similarity lower than a threshold indicates where the text should be segmented.

In this chapter, we evaluate the application of several topic modeling methods to two conversation corpora. We evaluate each topic model for three different representations of the datasets: where each document is either the concatenated conversation, i.e. on a conversation-level, each document is a single utterance from a conversation, i.e. on a utterance-level, or each document is a segment of several utterances from a conversation, i.e. on a segment-level.

As a baseline for comparing short and conventional length topic modeling, we include Dirichlet Multinomial Mixture model (GS-DMM) [50] and LDA, respectively. LDA sees the most frequent use when topic modeling for conversation is considered. However, we believe contextual information can be of benefit, especially when applied to utterance level data. Therefore, we also include three models using sentence embedding information, namely BERTopic [156], TopClus [160] and Combined Topic Model (CombinedTM) [138].

Furthermore, we compare the performance of sentence embedding trained on supervised data from a different domain, to unsupervised sentence embedding methods on the counseling corpus domain. We show that the contextual models perform better than both non contextual methods, and that BERTopic using utterance-level and segment-level data performs the best on the majority of evaluation metrics.

4.2 Background

4.2.1 Topic modeling in conversation

LDA is a generative topic model, where a document is seen as a mixture of topics and a topic is seen as a distribution over words. These topic mixtures are then sampled from a Dirichlet distribution. LDA is a basic but often powerful enough method for the goal of topic modeling. LDA has been frequently used for the purpose of basic conversation topic modeling. Dinakar et al. [52] used a variant of LDA in a crisis line, supported by expert input, to create interpretable and coherent topics. Wang et al. [70] proposed a model to discover topics from healthcare chat logs. Their method adapted LDA to capture noise in the form of latent personal interests of chat users.

4.2.2 Short texts

When topic modeling is applied to short texts like messages, the data becomes very sparse. This sparsity results in problems when models rely on document-level word co-occurrence. To tackle the sparsity problem, many different approaches have been considered. Short texts can be heuristically aggregated into longer pseudo documents as a straightforward solution [30, 31]. However, sometimes this is not possible or desirable. Other methods are based on making the stronger assumption that a document covers only one topic. Biterm Topic Modeling (BTM) [44] is a popular method of topic modeling for short texts. Instead of word-document co-occurrence, BTM models the co-occurrence of word pairs (bi-terms) in the entire corpus. This method has also been extended with the use of word embeddings [88]. To deal with the problem of noise in short text, the so-called Common Semantics Topic Model implores a common topic to which it assigns noise words [87]. Rashid et al. approach the sparsity problem using a fuzzy clustering approach named Fuzzy Topic Modeling (FTM) [107]. In fuzzy clustering, words can belong to multiple clusters based on a membership function. FTM first applies dimensionality reduction through PCA and afterward uses fuzzy c-means clustering to assign words to topics. Yin et al. propose a short text clustering method called Gibbs Sampling for GSDMM [50]. This generative method assumes a document corresponds to a single topic. Unlike LDA, documents are generated using the same topic.

4.2.3 Topic modeling using variational inference

Several methods have been proposed to incorporate neural networks for the field of topic modeling, the most successful of which are based on variational inference. ProdLDA is a method that uses variational autoencoding as a neural network approach to infer the LDA posteriors [81]. Word embeddings obtained through methods such as continuous bag-of-words by Mikolov et al. [42] are also used in topic modeling. Using word embeddings, Dieng et al. [119] developed the Embedded Topic Model (ETM). ETM is a generative model that, like LDA, models a document as a mixture of topics. However, ETM uses distributed representations for both words and topics. The topic embeddings are inferred using variational inference. A decoder network reconstructs the words belonging to the topics. Based on this method, CombinedTM [138] concatenates sentence embeddings from SBERT to the bag of words representation of a document and uses the concatenation as input for the autoencoding model of ProdLDA. Using these embeddings CombinedTM shows improved performance over ETM and ProdLDA.

4.2.4 Topic modeling using pre-trained embedding

Top2Vec [114] is a topic modeling method uses clustering of document embeddings, where the resulting clusters form the basis for the topics. After embedding the documents, dimensionality reduction is applied to the resulting embeddings using Uniform Manifold Approximation and Projection for Dimensionality Reduction (UMAP) [123]. UMAP reduces dimensionality by approximating a higher-dimensional manifold and projecting it to a lower dimension. Afterward, HDBScan is used to cluster the resulting transformed data. HDBScan is an extension of the density-based spatial clustering of applications with noise (DBScan) algorithm [38]. Using hierarchical clustering, it can accurately detect clusters of varying densities and shapes, while avoiding noise. Furthermore, HDBScan requires little in terms of hyperparameter optimization. BERTopic [156] adapts this method, to use document embeddings obtained through Sentence-BERT (SBERT) [108], a modification of BERT using siamese networks. BERTopic combines the tokens of the documents assigned to a cluster into a single set. For each set, words with the highest TF-IDF value are used to describe the topic represented by that cluster. Sia et al. [131] used an approach similar to Top2Vec. They applied PCA dimensionality reduction along with K-means and Gaussian Mixture Models on word embeddings. Additionally they used weighted clustering and term reranking to obtain the topics. A weakly supervised approach was proposed by Meng et al. [124] where category names can be provided to increase the interpretability of the topic models. TopClus [160] uses pretrained models to learn topic representations from a latent spherical space. The spherical space benefits clustering of the embeddings while also reducing the dimensionality.

4.3 Methods

This section covers the main steps for evaluating the different models. First, we highlight which models we chose to evaluate. Second, we explain which conversation corpora we used. Third, we cover the text pre-processing steps. Fourth, we show the different evaluation metrics used.

4.3.1 Models

We chose several topic models to evaluate on the utterance- and conversation-level data. As a baseline for short and conventional length topic modeling, we used GSDMM and LDA, respectively. For embedding-based models, we use BERTopic, TopClus and CombinedTM. Lastly, we adapted BERTopic, where instead of the BERT embeddings with averaged word2vec embeddings. We will refer to this variant as Clustering W2V. This allowed us to observe the impact of BERT embeddings had compared to a less computationally expensive embedding method. Every topic model method was tested on both utterances and the full conversation. Each topic model was optimized for 30, 60, and 90 topics. Hyperparameters for each model were optimized and the models with the best topic coherence scores were reported.

4.3.2 Datasets

For evaluation of the models, we used three corpora. The first corpus is a counseling corpus from a suicide prevention helpline. This corpus is difficult to model with traditional bag-of-words

Table 4.1: Utterance and conversation frequencies

Corpus	Utterances	Conversations
Counseling corpus	55,811	4,508
AnnoMI	9661	133
MultiWoz	72,022	7,679

methods such as LDA. The second corpus we included is an English mental health dataset, of transcribed interviews using the motivational interviewing paradigm called AnnoMI [168]. The third corpus we included is the Multi-Domain Wizard-of-Oz dataset (MultiWoz). MultiWoz is a dataset of dialogues containing conversations covering topics from multiple domains. We included the MultiWoz corpus to highlight the difference between the more frequent topic modeling setting existing in literature and the mental health corpora. MultiWoz covers specific domains where local context is less important, whereas this is not the case for the other two corpora. As mentioned previously, models will be tested both on utterances and full conversations. This results in a total of six datasets. Table 4.1 shows the sizes of each dataset.

The counseling corpus dataset contained chat conversations from a suicide prevention helpline. In this corpus, a help seeker contacts a counselor with an issue regarding suicide. It was the counselor's job to listen to a help seeker and to explore options with this help seeker where necessary. This corpus contained conversations covering many topics. Conversations also contained interactions that were not part of recurring topics. The MultiWoz dataset is a conversation dataset spanning multiple topics and domains. Compared to the counseling corpus, MultiWoz contains shorter conversations and covers strict topics, with less interference from small talk for example. Because of this, MultiWoz was expected to have less variability in the performance.

4.3.3 Document representations

Topic modeling uses documents as input, where each document can potentially contain a number of topics. In this study we looked at three ways to represent single document in the context of chat conversations, and how this impacts several topic modeling methods. First we represented an entire conversation as a single document (conversation-level), by concatenating all utterances. In the context of chat data, we defined an utterance as anything a participant says until responded to by another participant. Second we represented a single utterance as a single document (utterance-level). Third, we segmented a conversation into groups of consecutive utterances and used the concatenation of a group as a single document (segment-level).

The segment level data was created using the following four steps:

- We trained a binary classifier on a RoBERTa network with a next sentence prediction (NSP) objective. We used pairs of consecutive messages and pairs of random messages from the datasets as the training data.
- Using the NSP network, consecutive messages pairs were scored.
- Message pairs that scored below a threshold were marked as the end and beginning of two sections. This threshold was set such that segments averaged five messages.
- Utterances belonging to the same segment were concatenated.

4.3.4 Pre-processing

Pre-processing of the text consisted of six steps. First, all non-alphabetic characters were removed. Second, all text was lower-cased. Third, the text was lemmatized. Fourth, stop words were removed using the NLTK library of stop words. Fifth, the text was tokenized, removing any tokens that were shorter than three characters. Sixth, all but the 2,000 most frequent remaining tokens in the corpus were filtered out. For the sentence embedding, the text was only cleaned minimally, by removing special characters. On the utterance datasets, utterances with fewer than five words were removed. The chat messages and tokens were aggregated for each conversation to create the conversation dataset.

4.3.5 Pre-trained sentence embedding

For the sake of consistency, we used the same SBERT model for the topic modeling methods that leveraged sentence embedding. The models we used were "paraphrase-mpnet-base-v2" and "paraphrase-multilingual-mpnet-base-v2" for English and Dutch texts, respectively. The input for sentence embedding was minimally preprocessed. However, preprocessing was done to obtain tokens to describe each topic after clustering.

4.3.6 Unsupervised fine-tuning

It is worth noting that the domain of helpline conversations differs from the training data used for pre-trained networks. In addition to employing sentence embedding techniques via pre-trained networks, our study explored an alternative method for creating embeddings.

State-of-the-art sentence embedding models predominantly rely on supervised training, which involves labeled data comprising sentence pairs and their corresponding similarity scores. However, this is not always available, as is the case for the counseling corpus we are exploring in this study. To address this limitation, we implemented two unsupervised fine-tuning approaches and compared them to the outcomes using the pre-trained embeddings.

First, we used a pre-trained Dutch RoBERTa model and fine-tuned it using a triplet loss function. The triplet loss function, in this context, is defined as follows:

$$\text{Loss}(A, P, N) = \max(d(A, P) - d(A, N) + \text{margin}, 0)$$

Where d is some distance function. A, P and N are embeddings for an anchor, a positive and a negative message, respectively. The positive message is of the same class as the anchor message while the negative message is of a different class. This approach encourages the network to put at least a *margin* of distance between the anchor-positive pair and the anchor-negative pair.

To apply this to the counseling conversation dataset, we selected a pair of consecutive messages from a given conversation as the anchor and positive samples. Subsequently, we chose a message from a different sentence pair as the negative sample. This approach assumes that consecutive messages are more likely to be related, whereas two randomly selected messages are more likely not to be related. The training dataset consisted of all possible positive pairs. The negative message was randomly selected from the same batch. We used mean pooling to obtain the sentence embeddings, and fine-tuned the RobBERT network, a RoBERTa network trained on a large Dutch language corpus.

Second, we used the Tranformer-based Denoising AutoEncoder (TSDAE) to fine-tune the same pre-trained Dutch RoBERTa model [150]. Instead of generating data through the conversation structure, TSDAE introduces noise to the data and trains an autoencoder to denoise this data. The embedding of the class token represents the sentence embedding.

We compared these two unsupervised approaches to the results of the supervised pre-trained sentence embedding model for the BERTopic model.

4.3.7 Evaluation metrics

We used two metrics to evaluate the models. First, we computed the topic coherence using the C_v metric [59]. This metric combines normalized pointwise mutual information and cosine similarity with a sliding window. To keep the comparison unbiased, topic coherence was calculated for the full conversation variant of each corpus. Second, we used Word2Vec embeddings for an indication of semantic relatedness. For each topic, we computed the average pairwise cosine similarity of the word embeddings using the top-5 words of the topic. We discount this within-topic relatedness by the between-topic relatedness, using the inverse of the average pairwise cosine similarity between the average Word2Vec embedding of each topic. We define the word embedding score $W = W_{\text{within}} W_{\text{between}}^{-1}$

where

$$W_{\text{within}} = \frac{1}{kl(l-1)/2} \sum_{i}^{k} \sum_{j}^{l} \sum_{m=j+1}^{l} sim(w_{ij}, w_{im}) \tag{4.1}$$

$$W_{\text{between}} = \frac{1}{k(k-1)/2} \sum_{i}^{k} \sum_{j=i+1}^{k} sim\left(\frac{\sum_{m}^{l} w_{im}}{l}, \frac{\sum_{m}^{l} w_{jm}}{l} \right) \tag{4.2}$$

We have w_{ij} as the word embedding for word j in topic i. The number of topics is denoted by k where $k = 30, 60, 90$. The number of selected words per topic is denoted by l. In our evaluation we let $l = 5$, using the top five best ranking words for each topic, according to each topic model's own metric of ranking.

4.4 Results

4.4.1 Topic coherence

Table 4.2 shows the topic coherence for all models on the counseling conversation corpus. Here we can see that BERTopic scores higher than the other models on all evaluated topic sizes. We observe for 90 topics that the segmented conversations perform better, but the smaller topic sizes show better performance using only the utterances. Notably, the sentence embedding-based CombinedTM and TopClus improved using utterances or segments, instead of a full conversation, most likely due to the length limit of the sentence embedding model. As expected, LDA decreased in performance due to the sparsity of the smaller documents. BERTopic with average word embeddings on utterances also performed well across all topic ranges. GSDMM was not able to produce enough topics on the full conversation dataset and was left out of the results. On the full conversation corpus, LDA performed similarly to most other models.

Table 4.3 shows the coherence scores for the AnnoMI corpus. Models trained on utterances and segmented also do better than on the full conversation for this dataset. However, LDA and GSDMM

Table 4.2: Topic Coherence for counseling conversation corpus

Models		Number of topics		
		30	60	90
Utterances	LDA	0.45	0.45	0.45
	GSDMM	0.50	0.51	0.50
	CombinedTM	0.52	0.52	0.52
	BERTopic	**0.60**	**0.61**	0.59
	Clustering W2V	0.54	0.56	0.55
	TopClus	0.53	0.52	0.53
Segmented	LDA	0.47	0.47	0.47
	GSDMM	0.53	0.54	0.53
	CombinedTM	0.56	0.46	0.46
	BERTopic	0.53	0.59	**0.62**
	Clustering W2V	0.54	0.53	0.57
	TopClus	0.55	0.50	0.53
Full conversation	LDA	0.52	0.52	0.52
	GSDMM	-	-	-
	CombinedTM	0.46	0.46	0.46
	BERTopic	0.51	0.52	0.51
	Clustering W2V	0.50	0.52	0.52
	TopClus	0.45	0.45	0.46

perform much better on the AnnoMI compared to the counseling conversations. BERTopic shows the best performance for 30 and 60 topics and LDA show the best performance for 90 topics.

Table 4.4 contains the coherence scores for the MultiWoz corpus. On this corpus, BERTopic also outperformed the other algorithms that were tested. However, this time the full conversation performed better than the utterances. Notably, all the clustering methods on both utterances and conversation performed better than the generative methods.

4.4.2 Average pairwise word embedding similarity

The word embedding similarity scores can be found in Table 4.5. For this metric, the differences are more pronounced compared to the topic coherence. On the counseling conversation corpus, we again find BERTopic to outperform other models using utterance data. Similarly, BERTopic with word2vec embeddings also show good performance. It also performs better than both sentence embedding methods on the full corpus. This is most likely due to the length limitation of sentence embedding. Low scores for LDA

Table 4.3: Topic Coherence for AnnoMI

Models		Number of topics		
		30	60	90
Utterances	LDA	0.59	0.59	**0.60**
	GSDMM	0.57	0.60	0.59
	CombinedTM	0.45	0.46	0.44
	BERTopic	0.59	0.59	0.57
	Clustering W2V	0.58	0.57	0.57
	TopClus	0.50	0.51	0.50
Segmented	LDA	0.53	0.54	0.52
	GSDMM	0.57	0.60	0.59
	CombinedTM	0.42	0.43	0.44
	BERTopic	**0.59**	**0.61**	0.59
	Clustering W2V	0.58	0.59	0.57
	TopClus	0.47	0.46	0.47
Full conversation	LDA	0.48	0.48	0.48
	GSDMM	0.48	0.48	0.48
	CombinedTM	0.41	0.38	0.35
	BERTopic	0.50	0.49	0.50
	Clustering W2V	0.52	0.51	0.49
	TopClus	0.40	0.43	0.43

Table 4.4: Topic Coherence for counseling MultiWoz corpus

Models		Number of topics		
		30	60	90
Utterances	LDA	0.42	0.43	0.43
	GSDMM	0.56	0.57	0.57
	CombinedTM	0.63	0.60	0.61
	BERTopic	0.72	0.72	0.67
	Clustering W2V	0.72	0.70	0.65
	TopClus	0.65	0.65	0.63
Segmented	LDA	0.49	0.49	0.49
	GSDMM	0.47	0.47	0.48
	CombinedTM	0.60	0.60	0.58
	BERTopic	0.68	0.75	0.77
	Clustering W2V	0.71	0.69	0.66
	TopClus	0.67	0.69	0.65
Full conversation	LDA	0.57	0.60	0.59
	GSDMM	0.54	0.54	0.53
	CombinedTM	0.62	0.62	0.61
	BERTopic	**0.80**	**0.81**	**0.80**
	Clustering W2V	0.75	0.72	0.70
	TopClus	0.70	0.71	0.69

Table 4.5. Word embedding similarity scores for counseling conversation corpus

Models		Number of topics		
		30	60	90
Utterances	LDA	0.47	0.37	0.37
	GSDMM	0.45	0.46	0.50
	CombinedTM	0.78	0.73	0.78
	BERTopic	**1.70**	**1.87**	**1.77**
	Clustering W2V	1.28	1.44	1.37
	TopClus	0.67	0.66	0.66
Segmented	LDA	0.33	0.34	0.34
	GSDMM	0.45	0.45	0.46
	CombinedTM	0.70	0.66	0.65
	BERTopic	0.77	0.79	0.78
	Clustering W2V	0.80	0.81	0.80
	TopClus	0.65	0.59	0.63
Full conversation	LDA	0.52	0.39	0.38
	GSDMM	-	-	-
	CombinedTM	0.65	0.66	0.65
	BERTopic	0.52	0.58	0.59
	Clustering W2V	0.89	0.80	0.81
	TopClus	0.67	0.66	0.66

demonstrate the difficulty this method has with this type of noisy data. Even on the full conversation corpus, LDA does not perform well. Furthermore, most models perform better when using the utterance datasets.

Word embedding scores for the AnnoMI dataset can be found in Table 4.3. LDA on full conversation showed overall the best scores. This is in contrast to the poor performance for topic coherence LDA obtained. For the other methods, utterance and segmented approaches performed better.

Table 4.7 contains the word embedding scores for MultiWoz. The highest scores are seen in the models using utterance data, shared between the clustering models.

4.4.3 Pre-trained and unsupervised sentence embedding

Table 4.8 shows the topic coherence using BERTopic with pre-trained embeddings and fine-tuned embeddings using the triplet loss and TSDAE unsupervised methods. Both fine-tuning methods show a marginal improvement over only pre-trained embeddings.

Table 4.6: Word embedding similarity scores for AnnoMI

Models		**Number of topics**		
		30	60	90
Utterances	LDA	0.76	0.77	0.77
	GSDMM	0.80	0.79	0.79
	CombinedTM	0.58	0.60	0.55
	BERTopic	0.76	0.75	0.72
	Clustering W2V	0.75	0.76	0.70
	TopClus	0.59	0.61	0.58
Segmented	LDA	0.79	0.76	0.77
	GSDMM	0.75	0.76	0.78
	CombinedTM	0.59	0.58	0.55
	BERTopic	0.70	0.70	0.69
	Clustering W2V	0.71	0.69	0.70
	TopClus	0.59	0.57	0.60
Full conversation	LDA	**0.83**	**0.82**	**0.82**
	GSDMM	-	-	-
	CombinedTM	0.57	0.56	0.55
	BERTopic	0.75	0.78	0.76
	Clustering W2V	0.69	0.71	0.68
	TopClus	0.62	0.64	0.61

Table 4.7: Word embedding similarity scores for MultiWoz corpus

Models		**Number of topics**		
		30	60	90
Utterances	LDA	0.40	0.26	0.26
	GSDMM	0.23	0.23	0.25
	CombinedTM	0.53	0.51	0.50
	BERTopic	0.52	**0.62**	0.60
	Clustering W2V	**0.58**	0.52	**0.66**
	TopClus	0.54	0.54	0.51
Segmented	LDA	0.42	0.38	0.31
	GSDMM	0.24	0.20	0.27
	CombinedTM	0.36	0.37	0.46
	BERTopic	0.50	0.52	0.59
	Clustering W2V	0.53	0.48	0.55
	TopClus	0.53	0.54	0.54
Full conversation	LDA	0.28	0.33	0.34
	GSDMM	0.19	0.19	0.17
	CombinedTM	0.32	0.34	0.31
	BERTopic	0.44	0.45	0.44
	Clustering W2V	0.36	0.43	0.45
	TopClus	0.29	0.34	0.31

Table 4.8: Topic Coherence for different sentence embedding methods using BERTopic on the counseling conversation corpus

Models	Number of topics		
	30	60	90
Pre-trained SBERT	0.60	0.61	0.59
Tripletloss	0.61	0.61	0.60
TSDAE	0.60	0.61	0.61

Table 4.9: Top 5 topic words for BERTopic on the counseling corpus

Topic	Word 1	Word 2	Word 3	Word 4	Word
1	parents	mother	my	and	you
2	to sleep	tired	sleep	bed	me
3	friends	who	friend	you	with
4	music	listening	hearing	to	the
5	at home	lonely	room	alone	house
6	mindfulness	search	a	practice	breathing exercise
7	medication	pills	me	drinking	have
8	wound	blood	bleeding	knife	care
9	writing	reading	book	me	what
10	url	link	site	website	urge
11	safe	keep	safety	yourself	you
12	watch	series	film	netflix	youtube
13	eating	dinner	eating dissorder	me	cooking
14	sports	gaming	excercise	games	fifa
15	thoughts	my	suicide	suicide thoughts	that

Since the difference is small, the additional fine-tuning of the dataset could be omitted.

4.4.4 Topic words

Table 4.9 the top-5 topic words for fifteen topics using the best performing BERTopic model. Clear topics can be discerned where this was not possible using classical LDA methods on either utterance- or conversation-level data.

4.5 Discussion

This study explores topic modeling in mental health conversations, focusing on the challenge posed by the lack of words with descriptive power. We hypothesized that utilizing the full text of the conversation might be more challenging due to this limitation. Additionally, mental health conversation corpora may exhibit topics occurring very frequently across many documents, as well as topics occurring very infrequently.

To address these challenges, we proposed that dividing documents into smaller sections could enhance the coherency of topic modeling. Our analysis, based on the results, revealed that BERTopic outperformed other methods for most of the different corpora and topic sizes. Notably, the highest performance was consistently observed in the utterance and segmented datasets.

Furthermore, we see that methods using HDBScan performed well. The property of HDBScan to deal with noisy segments, potentially contributed to increased flexibility in modeling conversations. The results not only supported our initial hypotheses but also indicated that hierarchical clustering yielded the best performance. This outcome aligns with the observation that topics in mental health conversations have great variance in their occurrence.

The exception was the coherence scores for the MultiWoz dataset. We believe that this might be due to the length and general topic amount being lower for MultiWoz than they are for the counseling corpus. On the word embedding similarity metric, all models except LDA improved in performance when using utterances instead of the full conversation. The downside of this method is that relationships between utterances belonging to the same conversation are not considered. Therefore, a topic can only be as specific as can be expressed in a single message. However, we also found that ignoring noisy messages leads to better topic coherence and word similarity.

Between the utterances datasets and the segmented datasets, we observed that both obtained the highest topic coherence on multiple occasions. However, for the word embedding similarity, the utterance dataset outperformed the segmented dataset.

Our study suggests several opportunities for future research. An important limitation of this study is the limited generalizability, and including multiple helplines would help in this respect. While we did include two mental health datasets, the AnnoMI dataset is not as extensive as the suicide counseling dataset.

To address the sparsity constraints of short text, while considering noisy utterances. A possible solution is to use hierarchically constructed features from both utterance and conversation level data. Generative neural models such as ProdLDA and CombinedTM function through variational autoencoders [81, 138]. Hierarchical autoencoders have also shown good performance on several tasks [149]. This could extend the variational autoencoder for topic modeling to take local context into account. A variant of this has been proposed by [56] using hierarchical LSTM models.

4.6 Conclusion

Topic modeling can be difficult on datasets like conversation data, where local context, emotion and subtext is important. However, by reducing granularity, a local context can be incorporated using sentence embedding. We found that clustering of sentence embedding with noise using BERTopic results in more coherent topic models for conversation data when compared to other topic modeling methods. For the domain of topic modeling in conversation, we saw that BERTopic over utterances outperforms other models on conversation in both counseling corpus as well as the easier to model. Furthermore, based on the inspection of the topic models, we found the topics to be the most interpretable. This method is particularly useful when the overall topics within a corpus are of interest. Finally, we believe this can especially be of use for mental health services to gain insight into their conversations.

Chapter 5

Detecting topic changes in helpline conversations during COVID-19

Salmi, S., Mérelle, S. Y. M., Gilissen, R., van der Mei, R. D. & Bhulai, S.

BMC Public Health, 22 (Mar. 2022)

Abstract

Background
Preventative measures to combat COVID-19 have led to social isolation, loneliness, and financial stress. This study aims to determine if the pandemic is related to changes in suicide-related issues among users of a suicide prevention helpline.

Methods
A retrospective cohort study analyzed chat data from 113 using BERTopic, a natural language processing method. Messages from December 1, 2019, to June 1, 2020 (N = 8589) were examined. Relative topic occurrence was compared before and during the lockdown, starting March 23, 2020, and analyzed by gender, age, and living situation.

Results
The topic of COVID-19 saw an 808% increase after the lockdown. There was a 15% increase in help seekers thanking counselors, with male and young users expressing gratitude more frequently (+45% and +32%, respectively). Mentions of coping methods like watching TV (-21%) and listening to music (-15%) decreased, as did mentions of suicide plans (-9%) and specific suicide locations (-15%). However, suicide plans increased among those over 30 (+11%) and those living alone (+52%). Male help seekers discussed emergency care (+43%) and anxiety (+24%) more, while those under 30 mentioned negative emotions (+22%) and low self-confidence (+15%) more often. Substance abuse mentions rose by 9% among those over 30.

Conclusion
While mentions of distraction and suicide plans decreased, expressions of gratitude for the helpline increased, underscoring its importance during the lockdown. Help seekers under 30, males, and those living alone exhibited more negative changes related to suicidality and should be closely monitored.

5.1 Introduction

The spread of the novel coronavirus (COVID-19) has put the lives in danger of many people around the world [134, 136]. Officially declared a pandemic, the measures to contain the spread caused many businesses to lose their income and the economy became uncertain. These problems emerged quickly after the virus took over. However, there is another, less obvious problem that the pandemic created.

For many people, the virus causes distress; for some, this can take the form of impaired mental health. In the months after the discovery of the virus on 27th December 2019, studies already found a decline in public mental health [132]. More importantly, the virus and the measures taken to contain it may affect risk of suicide [69, 75]. Scientists have expressed the fear that the number of suicides would rise, as government policy and measures to contain it would result in unemployment, financial stress, social isolation and loneliness [120]. O'Conner et al. observed an increase in suicidal ideation during the initial weeks of the lockdown in the UK [125]. Fortunately, the number of suicides has not increased in the first months of the pandemic [139, 143]. However, this does not provide a complete picture of COVID-19's impact. We need to remain vigilant.

In what way does COVID-19 affect people with suicidal ideation? The virus introduced financial uncertainty, and social distancing became the new norm. People no longer had easy access to their social or professional outlets, education was disrupted, and there was an increase in bereavement and traumas. With a lack of social interaction, adolescents are particularly vulnerable because they are in an important period for social and emotional development [117, 146]. Did any of the pandemic's changes contribute to a change in the problems that people with suicidal behavior or ideation experience?

The current study explores the impact of the COVID-19 pandemic in a specific at-risk population, i.e. help seekers that contact a suicide prevention helpline in the Netherlands (113 Suicide prevention). Suicide prevention helplines provide an accessible means of obtaining information on people who are struggling with suicidality. These helplines exist to counsel people through this struggle. These help seekers with suicidal ideation can contact the anonymous helpline by phone or chat. The chat data from these helplines can be used in combination with natural language-processing techniques to gain insights into the problems of people with suicidal thoughts during the pandemic.

Several studies have developed algorithms to detect changes in text content. Blei and Lafferty, as well as Wang et al. presented methods that model topic changes over time [18, 60]. In the domain of suicide prevention, Kumar et al. used topic modeling to detect changes in suicide-related content found in social media following celebrity suicides to measure their impact [55]. To the best of our knowledge, no study has been conducted to apply state-of-the-art NLP within the context of suicide prevention helplines.

With more than 100,000 conversations with help seekers in 2020, the chat service of 113 presents an opportunity: these conversations can be analyzed during the COVID-19 pandemic in comparison with the conversations before. The potential changes can be observed from the logs of the chat service.

This study therefore aims to identify whether the COVID-19 pandemic is related to changes in suicide-related problems for help seekers on a suicide prevention helpline. A second aim is to detect changes in specific risk groups, e.g. male and young help seekers and help seekers alone. By doing so, the current study may inform suicide prevention strategies for organizations in public and mental health that help people with suicidal ideation and inform other national and international helplines about the application of this innovative methodology.

To this end, this chapter intends to answer the following questions:

- What are the most common conversation topics of help seekers on the helpline?
- Has the occurrence of conversation topics changed after COVID-19 containment measures were introduced?
- Do different demographics show different changes in topic occurrence?

When it comes to different demographics and how they are affected, we hypothesize the following: The first hypothesis is that the observed changes in topic usage will differ between men and women, due to men being more sensitive to social-economic change [29]. Furthermore, we also suspect different outcomes for older and younger help seekers due to the difference in social economic impact the government measures. The second hypothesis is that the observed changes in topic usage will differ between people younger than 30 and older than 30, since adolescents particularly are at risk and even before the pandemic suicide rates were rising among young people in certain high-income countries [125, 126]. Finally, due to the social isolation of a lockdown, the third hypothesis is that people who live alone will show different topic usage from people who don't live alone [128].

5.2 Methods

Due to the nature of helpline chat data, we conducted a retrospective cohort study. To answer the first question, topics were extracted from the chats, using a unique topic modeling method tailored to conversational data. To answer the second and third questions, conversation topics were compared before and after the introduction of the COVID-19 measures.

5.2.1 Sample

This study is centered around the first lockdown in the Netherlands, when the coronavirus first started to spread and impact the public. We believe that this period will show the greatest disparity between conversation topics. This lockdown took place on the March 23, 2020. Date spanned a roughly 6-month period, 3 months before and 3 months after the measures were announced (December 1, 2019–May 31, 2020). Conversations were used from December 1st onward to roughly match the number of conversations in both time periods and to include enough conversations from a period no information on COVID-19 was known.

The 6-month period yielded 8589 chat conversations that passed an initial triage stage. Situations where the help seeker was injured or could not continue the conversation in a safe environment were interrupted by the triage. These help seekers were subsequently referred to emergency services. These conversations were therefore excluded from this analysis. We used no further inclusion criteria.

For the purposes of this study, an anonymized database of this chat service was used. This introduced no research burden for the help seekers and therefore reduced the chance of selection bias. Of these conversations, 5179 (60%) and 3410 (40%) took place before and after March 23, respectively. Of the total sample, 1635 (19%) were male, 6576 (77%) were female, and 378 (4%) indicated to have a gender identity other than male or female. Furthermore, 6610 (80%) of the help seekers were under 30 years old, and 1662 (20%) were over 30 years old. Finally, 807 (9%) of the help seekers lived alone. There was a faulty age value in 300 conversations and these were declared missing values. The category 'other' for gender contained a small number of conversations. To avoid issues with the sample size of different topics this category was excluded from the analysis.

5.2.2 Data preprocessing

Chat transcripts are often noisy and unstructured. Unlike regular texts, a chat transcript is a conversation between two people. This causes it to evolve organically, usually leading to a wide variety of topics in a single document. Furthermore, chat transcripts often include slang, shorthand, and spelling mistakes. Therefore, text needed to be preprocessed so that the information could be structured.

The Natural Language Tool Kit [11] filtered out non-descriptive words based on a common stop word list. For the purpose of limiting the use of slang and shorthand, and also to eliminate spelling mistakes, the text was narrowed down to the 2000 most frequent words in the entire corpus.

For our study, we were interested in the conversation topics of the help seeker. We also knew that helpline counselors had not changed their protocols since the arrival of COVID-19. Therefore, the messages of the counselors were excluded. Messages of help seekers that were under three words were also excluded, to limit basic messages like greetings.

5.2.3 Analysis

The topic model method used in this study is a clustering of embeddings, based on the BERTopic method [156]. The embeddings were obtained using Sentence-BERT [108], a modification of the BERT network [98] to efficiently create short text embeddings. The full message was used for these embeddings. We used the pretrained multilingual BERT model “distiluse-base-multilingual-cased” to obtain the embeddings [129].

Using Uniform Manifold Approximation and Projection for dimensionality reduction, the embeddings are reduced from 300 to 5 dimensions [123]. Finally, the embeddings were clustered with HDBSCAN [77].

Separate embeddings were created for each message of the help seekers. Because the messages get clustered, each chat is a combination of clusters. This method mimics the approach of traditional topic modeling techniques like LDA [14]. BERTopic was preferred due to a poor fit of LDA when using the entire chat as input. Furthermore, due to the shorter length of the individual chat messages, we found the BERTopic method to provide better results over LDA, while needing less parameter tuning. This method was also compared to several other topic modeling methods, such as the contextual topic model [138], and found to provide the best fit on the data.

The clustering of chat messages resulted in 81 clusters. Some of the clusters were comprised exclusively of identical messages. For example, messages like "I don't know" were used by many help seekers and therefore identified as their own clusters. As these clusters did not contain any relevant information, they were excluded from the results. We identified fifteen clusters with this characteristic, resulting in 66 clusters from which the topics were derived.

Finally, three methods were used to obtain definitions for each of the topics. First, the preprocessed message contents were grouped according to their assigned cluster. The TF-IDF metric was used on the grouped text to generate a list of the top five words for each topic [156]. Second, we inspected a handful of the anonymized messages associated with each topic for extra context. Third, two experienced counselors also observed the top five words along with the anonymized messages for each topic and provided their definitions. Based on this information, the final decision on the topic definitions was made by the first and second author. They solved uncertainties and discrepancies by checking several anonymized messages from the topics and discussed additional insights. The first author selected these messages based on how central to the cluster these messages were.

Because the number of conversations between the two time periods was different, a relative measure was used to determine the difference in the use of a topic before and after the introduction of the COVID-19 measures. Let N_1 and N_2 denote the total messages for the first and second time periods respectively. Given n_1^t and n_2^t as the number of messages belonging to topic t in each time period, the change in topic occurrence becomes:

$$r^t = \frac{n_2^t}{N_2} / \frac{n_1^t}{N_1} - 1$$

Ultimately, this resulted in a rate of relative change for each topic, presented as a percentage in the results. These same methods were also used to compute the change in topic use for each of the different subgroups. To obtain the change for each topic among men, women, younger, older, and help seekers living alone, only the conversations that fell into that subgroup were included.

5.3 Results

5.3.1 Most important topics

Through the topic modeling of all 8,589 chat conversations 66 topics were determined. The definitions as well as the frequencies and relative change of these topics can be found in Table 5.1. The most frequent topics concerned events the help seeker experienced or treatment the help seeker received in the previous days or weeks. For example, help seekers talked about treatments they received or if they had discussed their suicidal thoughts with a specialist.

Table 5.1: Topic frequency and definition

Frequency	Relative change	Topic
6138	-1%	Description of events and treatment in recent days or weeks
5962	1%	Need to talk, difficulty talking to those around them
5076	-1%	Family circumstances
4964	3%	How the help seeker feels
4961	-3%	The help seeker doesn’t know what to do
4598	-1%	Thoughts and thought patterns
4511	-1%	Help seeker is seeking help / help isn’t working
4111	-1%	Life is pointless
3638	0%	Friendship
3476	-6%	Suicidal thoughts
2832	-2%	Desire to go to sleep or difficulty sleeping
2450	-5%	Help seeker, housemates, family are not at home
2439	-2%	Depression and stress
2288	5%	Psychologist or psychiatrist
2226	8%	Help seeker finds something difficult or awkward
2212	-1%	Help seeker has been trying things for a long time
1832	3%	Pain and sadness

Continued on next page

Table 5.1 – continued from previous page

Frequency	Relative Change	Topic
1791	-2%	Problems with medication and alcohol
1732	-20%	School and study
1706	-9%	Help seeker is looking for a solution or can't see a solution for their problems
1679	5%	Help seeker's hopes and expectations
1589	-9%	Has plans die by to commit suicide
1503	10%	Thanking the counselor
1463	-8%	Description of events and behaviors in the last month(s)
1305	1%	The help seeker has a lot on their mind and wants peace
1229	0%	Not being understood or not understanding others
1192	4%	No confidence in self or others
1182	-2%	Self-harm
1173	4%	Desire to be happy
1161	13%	Writing down thoughts, writing goodbye letters, and distraction by reading books
1142	5%	Question for the counselor
1139	4%	Panic and anxiety
1016	9%	Apologizing to counselor
970	11%	Desire to contact the helpline or someone else
887	-15%	Distraction by listening to music or reading
828	5%	Negative feelings or self-image and a desire for more positive thoughts
750	-21%	Distraction by watching TV, series or films
683	-2%	Crisis service
623	-14%	Planning to commit die by suicide at a specific location
620	2%	Help seeker can't go on, has given up
608	6%	Children
606	6%	Help seeker finds it difficult to talk about it
583	5%	The help seeker has no more options apart from committing dying by suicide
567	808%	Coronavirus
549	4%	The help seeker asks how he/she can carry on living
500	12%	The help seeker is in his/her room
479	-8%	The help seeker has doubts or is uncertain
472	1%	Help seeker has worries about the future
462	-9%	Help seeker wants to tell their story
459	-1%	The help seeker is doing better
427	-5%	The help seeker is bleeding
417	-7%	Help seeker has no more energy
412	13%	Help seeker indicates that a proposed solution doesn't work
412	-1%	Negative emotions and weeping
390	-5%	Feelings of guilt and responsibility
368	15%	Help seeker is grateful for being listened to
368	-9%	Means for of committing suicide in the kitchen
358	3%	Help seeker is grateful for the conversation
337	16%	The help seeker is experiencing something for the first time
322	1%	The help seeker is safe
319	31%	Help seeker is at home
308	1%	Distraction by taking the dog for a walk
301	5%	Loss of control

Continued on next page

Table 5.1 – continued from previous page

Frequency	Relative Change	Topic
273	-32%	Romantic relationships
271	18%	Involvement of police in the event of a suicide attempt or unsafe home situation
251	11%	Problems due to autism or ADHD

5.3.2 Changes in topic occurence

As expected, during the initial months of the pandemic, help seekers more often contacted from home and more often mentioned encounters with new situations. Conversations had more mentions of negative emotions. However, we also saw an increase in conversations that mentioned encounters with law enforcement, connected to suicide attempts or unsafe environments at home. Furthermore, problems with autism and ADHD appeared more often in chat conversations during the lockdown.

5.3.3 COVID-19 as a conversation topic

The most extreme result was, perhaps unsurprisingly, the topic of the COVID-19 pandemic, which saw an 808% increase in relative occurrence after the introduction of COVID-19 containment measures. Table 5.2 shows some of the common problems help seekers mentioned, as well as some example chat messages related to these problems.

5.3.4 Changes for gender, age, and living alone

Because relatively more women and young people contacted the chat service, their results are largely similar to the total results. Tables 5.3 and 5.4 show the most remarkable differences between the two gender and age demographics, respectively. Table 5.5 shows the biggest changes for help seekers who live alone.

After the introduction of COVID-19 measures, conversations with male help seekers more often involved feelings of panic, fear, and emergency care. They also expressed their gratitude more frequently.

When looking at changes between the different age groups after the introduction of COVID-19 measures, we saw that younger help

Table 5.2: Main categories within the COVID-19 topic

COVID-19 related problem	Example chat message
Loneliness	The loneliness has become worse because of corona.
No distractions / lack of structure	I also don't have a lot of distraction now because everything is closed due to corona.
Change in regular care	It seems to get more difficult in this corona period because ...
Feeling of getting locked up	... a fear getting stuck at home because of corona.
Fear of the virus	... with my fear of germs this whole corona situation is a nightmare come true.
Difficult situation at home	... and because of corona I work from home, it's so stressful, and also not the best with my family.
(Threat of) unemployment	I was fired from my other job because of corona.
Increase of fear and depression	Due to this whole corona situation my anxiety is flaring up.
Concerns about somebody else	And I'm not allowed to go near my grandparents because my granddad has ... and if he gets corona he will not survive
Alcohol and drug abuse	With corona it's worse than usual. I'm constantly worrying and then I drink ... or cut myself to cope.

seekers focused more on negative emotions. They also expressed a more frequent lack of confidence in themselves or others. Furthermore, the younger help seekers did more often express thankfulness for the conversation during the lockdown. On the other hand, older help seekers more often discussed problems with medication, alcohol, and methods of self-harm.

Help seekers that lived alone showed the most extreme changes between the two time periods. This is in part because this group was relatively small compared to the total number of help seekers. Nevertheless, there were striking changes in this group. Help seekers that lived alone more often talked about suicidal thoughts or plans, and also about having no energy. On the other hand, the conversations less often mentioned difficulty talking about their situation or about self-harm.

To summarize, there was an increase in topics indicating gratefulness for the helpline. Furthermore, there was a decrease in mention of plans for suicide and suicidal thoughts. Help seekers who live alone however, showed an increase on plans for suicide and suicidal thoughts. Male help seekers showed increase mention of panic and anxiety. Help seekers under 30 showed an increased mention of negative emotion and lack of self-confidence, while help seekers

Table 5.3: Comparison of topic change between men and women

Total	Women	Men	Topic
31%	44%	-15%	Help seeker is at home
12%	21%	-13%	The help seeker is in his/her room
6%	5%	-14%	Help seeker finds it difficult to talk about
4%	2%	24%	Panic and anxiety
4%	13%	-21%	Desire to be happy
3%	-2%	43%	Help seeker is grateful for the conversation
1%	5%	-26%	The help seeker has a lot on their mind and wants peace
-2%	-5%	43%	Crisis service

Table 5.4: Comparison of topic change between help seekers younger and older than 30

Total	Age < 30	Age >= 30	Topic
15%	31%	-12%	Help seeker is grateful for being listened to
13%	22%	-12%	Negative emotions and weeping
4%	15%	-21%	No confidence in self or others
4%	13%	-27%	Desire to be happy
3%	15%	-20%	Help seeker is grateful for the conversation
-2%	-5%	9%	Problems with medication and alcohol
-9%	-12%	11%	Means for of committing suicide in the kitchen

over 30 showed an increased mention of problems with medication or alcohol.

5.4 Discussion

This study was conducted to inform suicide prevention strategies for organizations in public and mental health that help people with suicidal ideation. This was also the first study in the field of suicide prevention that used BERT embeddings resulting in an in-depth analysis of a large number of conversations using topic modeling. This study differs from previous studies on mental health conversational data. Examples of such studies dealt with real-time topic modeling of crisis chat helplines with LDA that leveraged experts during the training to improve the model [52], and modeling of a set amount of conversation stages [62].

The topic model found multiple changes in topic frequency in the overall help seeker population. Several of these changes can be explained by the government measure. Examples of these

Table 5.5: Topic change of help seekers who live alone

Total	Living alone	Topic
-20%	15%	School and study
-9%	52%	Has plans to commit die by suicide
-6%	27%	Suicidal thoughts
-2%	-41%	Self harm
6%	-62%	Help seeker finds something difficult or awkward
1%	-40%	Help seeker is safe
-7%	70%	Help seeker has no more energy

are that help seekers called more often from home, shared less about romantic relationships, and talked less about their education. These changes are in line with the government measures causing schools to close and recommendations to stay at home to keep social interaction to a minimum.

Certain changes pertaining to the functioning of the helpline were of interest. There was increased mention of thankfulness on the part of the help seekers for the conversation during the lockdown, indicating a desire for contact. This is an important finding since loneliness and thwarted belongingness are strongly associated with suicidal ideation and behavior [91]. In a study on reporting the COVID-19 related problems mentioned by help seekers, van der Burgt et al. found interruption of regular care and loneliness to be the most frequently mentioned [179]. Helplines can provide support in times of loneliness that has been amplified by the lockdown.

There was an increase in mentions of negative emotions and lack of self-confidence in chats, especially among younger help seekers. The mental health of this group was tracked in the UK by Pierce et al. [127]. They found an increase in mental distress in the general population during the pandemic and specific groups, like adolescents, were affected more significantly. A study on the age group 8–18 found indication that internalizing problems of this group increased during the pandemic [153]. We know that this group is already at risk for calling a suicide prevention help line. As an important period in their life for social emotional development and with government measures still persisting it is important to provide and promote means for this group to socially interact on a regular basis to improve their.

There was a decrease in mentions of specific plans and methods for suicide. Furthermore, helplines saw an initial decrease in usage of the helpline and the observed number of suicides during the early months of COVID-19, and there was no significant increase[143]. However, financial uncertainty and increased risk factors due to social isolation would expect the opposite. Suicide remains a complex phenomenon and we cannot yet explain how these findings relate.

There was also an increase in mentions of panic and anxiety among male help seekers. This could be because of the image of social economic change [29].

Among help seekers who live alone, an increase in mentions of specific suicide plans was observed. This group of people were hit harder by the pandemic than most help seekers, and this was apparent in the discussed topics. A possible contribution could be the decrease of their daily social contact during a lockdown. Combined with no contact at home, it could lead to more time with their own thoughts. The ruminative process affects the transition of defeat leading to entrapment which can then turn into suicidal ideation according to the model of O'conner et al. [91]. It will be important for policy makers to keep a close eye on how their actions affects people who live alone and provide leniency where possible so that contact for this at-risk group could still be possible.

5.4.1 Limitations

While technology has progressed rapidly in recent years, natural language processing remains a randomized process. Outliers or "noise" in the data is therefore a factor to be aware of, and we acknowledge this as a limitation of this study. With this method, noise could manifest in three ways. First, the topic model does not classify every message. This feature has the benefit of reducing false positives within the topic model, based on the closeness to a cluster and the cluster size. However, this does also mean that some of the messages that could be considered relevant to the topic can fall outside the cluster. The second limitation stems from the cluster size. Topics that have a smaller size are more prone to fluctuation and therefore have lower statistical power. This effect is amplified when considering a smaller demographic, like male help seekers or help seekers who live alone. The third limitation concerns the accuracy of the BERT embeddings. While no model is perfect, the potential mismatch between the text data used for training the BERT model and the chat data is something to be aware of. The BERT model was trained on a concatenation of Wikipedia pages [129]. As our dataset includes mostly conversational data, there might be language usage specific to conversations that is therefore unseen by the BERT model. This could influence the number of messages classified as noise or wrongly assigned to a cluster.

5.4.2 Implications

Several implications for counselors in a suicide prevention helpline are recommended. Some of the key pillars that counselors were trained to look for in a conversation were short-term coping, long-term hope to live, reasons to live, and decreasing accessibility of the means for suicide. Counselors can direct the conversation to cover topics that are less often mentioned during the lockdown. For example, two of the three topics about short-term coping styles are less frequently reported by help seekers. Difficulty of finding distraction also appeared as an recurring topic in a study by van der Burgt et al. [179].

People who live alone have no household to infect and we recommend policy makers to treat this group accordingly, by allowing exceptions for those who need it, to work from the office for example. Negative emotions among young people could have an increased chance to go unnoticed due to the decreased social interaction. Allowing for social interaction outdoors, such as sporting events could be a possible way for younger people to be able to talk to others and express emotions they cannot at home. This is also important because it is part of forming their identity through interaction with peers [117, 146], which would also promote self-confidence.

5.4.3 Future work

There are three main avenues of future work. First, the current method can be applied to other time periods. As of the time of writing, there has already been a second and third wave of the virus, and a second, stricter lockdown has gone into effect in the Netherlands. With distinctive periods, characterized by the different stages of lockdown, we can infer a better picture of help seeker trends over time. Second, we advise organizations that provide support through online counseling that are interested in obtaining insights on their demographic to consider this method. Third, topic interpretability can be improved. One approach involves levering the attention mechanism underlying the BERT model. A summary could potentially add semantic context that a list of important words could not. One potential way to do this is with transformer-based models for summarization [94, 133].

5.5 Conclusion

This study offers new insights into a specific at-risk population for suicide by examining common problems that help seekers from a suicide prevention helpline experience. Although the number of suicides fortunately has not increased during the COVID-19 pandemic in the Netherlands in 2020, changes in frequencies of these problems during the lockdown were detected. Help seekers showed a decreased mention of topics involving distractions and plans or methods for suicide. Topics involving gratefulness for the conversation saw an increase in mention.

Furthermore, the model detected several changes in sub demographics of the help seekers. Among help seekers who live alone there was in increase in mention of plans for suicide. Male help seekers showed an increase mention of panic and anxiety. Help seekers under 30 showed an increase in mentions of negative emotions and lack of self-confidence.

It remains important, during this period of reduced social interaction, increased fear and uncertainty, to keep monitoring the changes over time, to allow for exceptions for younger, male or help seekers who live alone where possible, and to keep help and communication available.

Part III

Classification and insights

Chapter 6

Classification of counselor utterances for motivational interviewing

Pellemans, M., Salmi, S., Mérelle, S. Y. M., Janssen, W. C. & van der Mei, R. D.

Journal of Medical Internet Research, Accepted for publication.

Abstract

Background

The rise of computer science and AI offers great potential for developing evidence-based health interventions. MI is one such intervention effective in improving various health behaviors, though challenging to master. This chapter explores AI models' performance in classifying MI behavior and their feasibility as automated support tools in online mental health helplines.

Methods

A dataset of 253 MI counseling sessions from 113, coded with the MI-SCOPE codebook, was used to train and evaluate four machine learning models and one deep learning model to classify MI behaviors.

Results

The deep learning model BERTje outperformed all ML models, accurately predicting counselor behavior (accuracy = 0.72, AUC = 0.95, Cohen's kappa = 0.69). It differentiated MI congruent- and incongruent counselor behavior (AUC = 0.92, kappa = 0.65) and evocative and non-evocative language (AUC = 0.92, kappa = 0.66). For client behavior, the model achieved an accuracy of 0.70 (AUC = 0.89, kappa = 0.55). The model's interpretable predictions discerned client change- and sustain talk, counselor affirmations, and reflection types, facilitating valuable counselor feedback.

Conclusion

AI techniques can accurately classify MI behavior, highlighting their potential to enhance MI proficiency in online mental health helplines. With sufficient dataset size and training samples, these methods can be applied to other domains and languages, offering a scalable, cost-effective way to evaluate MI adherence, accelerate behavioral coding, and provide quick, objective therapist feedback.

6.1 Introduction

MI is a client-centered counseling style that helps individuals change their behavior by resolving ambivalence using non-directive conversation techniques. It has been found effective in improving a wide range of health-related behaviors [41], such as weight management [161], addictive behaviors [57], and promoting self-management in patients with chronic health conditions [79]. During MI, counselors use a specific set of conversation techniques, most notably open-ended questions, and reflections, to let clients voice their own arguments for a particular behavior change. They are encouraged to elaborate on these reasons. This way, clients reason themselves into changing their behavior, strengthening their intrinsic motivation for the behavior change and avoiding the often triggered defensive mechanisms when others argue for such a change. MI is a crucial anchor in guiding the counseling process during chat-based conversations at 113.

Despite the simplicity of its principles, MI can be a challenging skill to learn, and MI requires substantial expertise to apply effectively [164]. Earlier research has shown that counselors at 113 applied MI techniques consistently during chat conversations but could not strategically deploy MI techniques to elicit enough change talk from clients to change their behavior intrinsically [158]. Therefore, it becomes imperative to improve the proficiency level of counselors in applying MI techniques to conduct conversations more effectively. Especially since eliciting change talk from clients accounts for the effectiveness of MI [89].

One way to achieve this is through the automated evaluation of counselor responses to clients' expressed language utterances. By increasing their behavior awareness, counselors can significantly reduce cognitive effort and reflect on MI insights for education. Multiple validated proficiency measures exist for MI [121], and tools are already in development to measure treatment fidelity automatically [102, 164]. In the context of chat-based helplines, these tools can provide counselors with immediate feedback during ongoing chats, potentially improving the quality of the service. Chat-based helplines also present a unique opportunity for developing such treatment fidelity tools due to the availability of extensive databases of written conversations. Also, [41] found that MI is a robust intervention across patient characteristics, which gives these tools broad applicability in numerous health settings.

6.1.1 Enhancing MI effectiveness through AI

AI has made a significant impact in recent years in many fields, the field of clinical mental health being no exception. AI offers enormous potential to analyze large datasets through ML algorithms. By analyzing data from MI sessions, ML algorithms can identify successful and unsuccessful applications of MI concepts, supporting and training MI practitioners. Additionally, counselors can use an AI support tool to evaluate the quality of their sessions. These tools can help improve and assess counselors' MI proficiency cost-effectively and tailor additional training to their needs.

AI can also speed up the coding of MI sessions, making it easier to analyze and provide feedback *during* and *after* a counseling session. Providing counselors with ongoing feedback seemed especially important for learning MI [47]. Besides, more immediate feedback has a powerful impact on skill development compared to delayed feedback [102].

Although large amounts of data are typically required to train ML models to perform well on complex tasks such as capturing MI behavior, AI has developed techniques that perform well in domains with limited available data, providing insights into developing and improving evidence-based health interventions [159].

Since behavioral coding is often time-consuming, several studies have explored the automated annotation of MI transcripts in counseling sessions using ML techniques. [86] conducted experiments on automating the annotation of weight loss counseling sessions using the MI Sequential Code for Observing Process Exchanges (MI-SCOPE) codebook. They assessed various classification methods, incorporating linguistic, contextual, and semantic features based on

Linguistic Inquiry and Word Count (LIWC) [58]. Their experiments showed that a Support Vector Machine (SVM) model with these features achieved 75% accuracy in automatically annotating MI transcripts containing seventeen behavioral codes. [101] aimed to develop a classification model to automatically code clinical encounter transcripts about weight loss using the MI-SCOPE behavioral code scheme. Their SVM model achieved a 69.6% F1 score on seventeen classes. [67] introduced two ML models for automatically coding MI sessions. The researchers found that the best-performing ML model had a good or higher utterance-level agreement with human coders (Cohen's kappa > 0.60) for open and closed questions, affirmations, and giving information. However, there was a poor agreement for client change-talk, client sustain-talk, and therapist MI-congruent behaviors. [78] presented a model for predicting MI counselor behaviors in multiple medical settings. Their Support Vector Machine (SVM) classifier performed well for more frequently encountered behaviors (reflections and questions) using N-grams, syntactic, and semantic LIWC features [58]. However, the performance varied much per predicted class, also obtaining lower performance for emphasizing autonomy and affirmations. [148] compared the classification performance of client behaviors throughout MI psychotherapy sessions with students having alcohol-related problems using pre-trained embeddings and interpretable LIWC features. Their best-performing model (pre-trained RoBERTa) achieved an F1-score of 0.66 in a three-class classification. [164] developed a Technology Assisted Motivational Interviewing Coach (TAMI) incorporating ML models to deliver MI predictions for counseling sessions about tobacco cessation. Using a novel deep learning architecture combining a large fine-tuned language model and graph theory, the automated change talk/sustain talk/follow-neutral classifier achieved an accuracy of 0.74 and an F1-score of 0.75.

For a comprehensive overview of research papers using ML to classify MI behavior for assessing treatment fidelity, we refer to [137]. The application domain significantly varies, which also applies to the reporting of coding reliability estimates. Assessing treatment fidelity and reliability holds enormous relevance for evaluating study quality and the successful integration of MI into practice. A meta-analysis on the effect of MI on medication adherence found that interventions which examined fidelity and provided counselors with feedback on their fidelity were more effective than those that did not [64], indicating that a higher fidelity may lead to improved intervention outcomes. [84] highlighted that fidelity is often poorly measured and reported. Moreover, MI adherence and fidelity demonstrated considerable variation across different settings and application domains [74, 84, 85].

Despite the promising algorithm performances, predicting MI-congruent counselor behavior and eliciting client change talk was challenging [67, 78]. Besides, few studies adhered to best-practice ML guidelines. Although testing methods on unseen data is an essential measure of method performance in ML, only a small proportion of studies tested their methods on hold-out data. A hold-out subset provides a final estimate of the ML model's performance after it has been trained and validated. Similarly, [137] found that almost half of the studies in their review did not describe how they undertook data pre-processing.

For readers to assume that ML methods will generalize on future data, researchers must report these methodological processes clearly and transparently, including robust coding reliability and fidelity measures. Previous studies showed the feasibility of providing feedback to counselors via a support tool [140], consistently measuring fidelity and reporting Krippendorff's alpha estimates for inter-rater reliability.

6.1.2 The present study

This study aims to investigate the performance of AI models in classifying MI behavior and explore the feasibility of using these models in helplines as an automated support tool for counselors in clinical practice. We use a coded dataset of 253 chat-based MI counseling sessions conducted at the chat helpline of 113. We train and compare different AI algorithms to classify client- and counselor MI behavior based on language use to identify the most suitable model for the task.

The key contributions of this paper are as follows: First, to the best of our knowledge, this is the first research that combines AI and Motivational Interviewing with a focus on suicide prevention. Second, We aim to assist counselors in a suicide prevention helpline to overcome the practical challenges of eliciting change talk and enhancing awareness of conversation quality by providing feedback. (iii) Our AI approach is described in detail, adhering to the best practices in the field and establishing a benchmark for implementing similar techniques in various settings.

6.2 Methods

6.2.1 Dataset

This study used a coded dataset of 253 chat conversations (constituting 23,982 chat messages, 12,125 counselor messages, and 11,857 client messages) from chat-based MI counseling sessions conducted at 113 between July 2020 and January 2021. All chats were Dutch language chats and lasted at least twenty minutes. [158] described the exact data collection procedure.

6.2.2 Participants

Participants in the dataset contacted the 113 crisis chat service in the Netherlands between 08:30 AM and 10:30 PM. All clients who spoke Dutch, filled out both a pre-and post-chat questionnaire, and reported at least some suicidal ideation on the pre-chat questionnaire (score ≥ 1 on a 7-point scale) were eligible for participation in the study [158].

The ethics review committee of the VU University Medical Centre in Amsterdam reviewed and approved this study (2020.105). The national legislation and institutional requirements did not require written informed consent from the participants.

For analysis, we only used (cleaned) text of the chat messages without any personalized meta-information (including - but not limited to - age, gender, ethnicity, or clinical diagnosis). There was no collection procedure for other additional data *before*, *during*, or *after* a chat conversation.

6.2.3 Procedure

Measures

Practical instruments exist to understand the quality and effectiveness of applying MI in counseling conversations. Researchers coded the dataset with the MI-SCOPE coding instrument [19]. Researchers created this tool to explore the relationships between essential theoretical constructs of MI, the therapy process, and client outcomes. The focus is on analyzing the relationship between MI-specific interviewer behaviors and subsequent client behaviors within an MI session. The MI-SCOPE combines two successful coding systems: the MISC [16] and the commitment language coding system developed by Paul Amrhein [13].

The MI-SCOPE provides five indices of treatment integrity, including the percentage of MI-consistent responses, the relative amount of open questions, the proportion of complex reflections, the reflection-to-question ratio, and the proportion of change talk. [121] indicated that reliability estimates for the MI-SCOPE are generally fair to excellent.

While most studies have used the MISC only [137] or the well-validated but relatively short MITI, these instruments do not provide information on the amount of change-talk and sustain-talk expressed by the client, whereas the MI-SCOPE does. The MI-SCOPE thus covers more aspects of MI, incorporating both client and counselor behavior, and is more time-efficient [106]. Had the MITI or MISC been used instead of the MI-SCOPE, [158] would not have detected the insufficiency of MI effectiveness in eliciting client change talk. Research by [46] and [162] also emphasizes the importance of fidelity measures in MI. The studies suggest that therapist adherence to MI techniques can influence client engagement and outcomes, and high-fidelity counseling can improve intervention effectiveness.

Dataset coding and reliability

Researchers who coded the dataset [158] followed recommendations by [125] and described the exact coding procedure in their paper. Of the total number of counselor messages, [158] labeled 9,177 counselor messages with fine-grained MI behavioral codes and 2,948 chat messages with less fine-grained codes, indicating only MI congruency.

The coding process for all MI conversations lasted four months, from January 1, 2021, to May 12, 2021. The researchers used the qualitative data analysis tool *Atlas.ti 9* for the coding procedure and assessing reliability estimates. Inter-coder reliability was sufficient, as [158] reported a Krippendorffs alpha-binary of 0.82 for the percentage of MI consistent responses and 0.90 or higher for open questions, closed questions, and reflections. Generally, researchers consider an alpha-binary over 0.90 acceptable in all cases, while an alpha-binary ranging from 0.80 to 0.90 is deemed sufficient.

Code grouping

To predict counselor behavior congruent with MI, we partnered with a seasoned psychologist at 113 (listed as the fourth author) to group the annotated MI behavioral codes - as outlined in the MI-SCOPE coding manual [17] - considering the practical challenges within the counseling process.

We combined all closed questions, negative and neutral reflections, and positive reflections (simple and complex), yielding seventeen code groups for counselor language. For counselor language, we created two groups of the MI-SCOPE codes based on whether counselor language elicited client change talk (7,477 non-evocative messages; 1,700 evocative messages) and whether it was MI congruent (8,765 MI-congruent messages; 3,360 MI incongruent messages) (see Table 6.1). We excluded the labels *Raise Concern* and *Direct* from further

Table 6.1: MI code groups for counselor language and whether or not a code is assigned evocative and/or MI congruent

MI code group	Evocative	MI Congruent
Advise with Permission	×	✓
Advise without Permission	×	×
Affirm (Aff)	✓	✓
Closed Question	×	×
Confront (Con)	×	×
Emphasize Control (Econ)	✓	✓
Filler (Fill)	×	×
General Information (GI)	×	×
Open Question (OQ+)	✓	✓
Open Question (OQ−)	×	✓
Open Question (OQ0)	×	✓
Permission Seeking	×	✓
Reflection (+)	✓	✓
Reflection (0−)	×	✓
Self-Disclose (Sdis)	×	×
Structure (Str)	×	✓
Support (Sup)	×	✓

analysis due to their low occurrence in the dataset. We did not assign detailed labels to the four client codes (*Ask, Follow/Neutral, Change Talk, and Sustain Talk*). Initial data analysis revealed that only 18.52% of all counselor messages were evocative, and 70.33% were MI congruent.

6.2.4 Analytic strategy

We trained and evaluated four ML models and one deep-learning model to classify client- and counselor MI behavior based on language use. ML models benefit from human-extracted features, while deep learning models learn complex patterns without feature selection. Although deep learning models have better performance potential, these models require more data and have less interpretable reasoning. We further describe the feature selection process, the models, and how we addressed these limitations of the deep learning model.

Feature selection

Available features: In total, we extracted 5,850 features for each of the MI-coded chat messages. These features included high-level concepts such as topic, grammar, and sentiment, as well as low-

Table 6.2: Overview of all feature categories, descriptions and corresponding feature sets.

	Feature category (# features)	Description	Feature set
1	Bag of Words (2,000)	Word occurrences in a chat message.	1
2	TF-IDF (2,000)	Relative importance of word occurrences across all chat messages.	1
3	Textual features (27)	Capturing a variety of textual information such as message length and the number of question marks.	2
4	Word embeddings (300)	Representing words as vectors of numbers in high-dimensional space to capture their semantic and contextual meaning.	3
5	Parts Of Speech (36)	Grammatical categories such as verbs, nouns, and prepositions.	4
6	Named Entities (18)	Real-world object categories (e.g., *Person*, *Location*, *Date*).	4
7	Dependencies (1,056)	Capture the grammatical structure of sentences by identifying relationships between the words.	4
8	Topics (42)	Identifying recurrent themes or topics.	5
9	Sentiment (29)	Extract emotions, appraisals, and attitudes toward different entities.	6
10	Cognitive Distortion Schemata (279)	Extracting language that indicates cognitive distortions (exaggerated or irrational thought patterns).	7
11	Temporal Patterns (63)	Capture the sequential message structure based on a temporal pattern mining algorithm.	8

level concepts such as counting the occurrence of each word. For a complete list of all feature categories used for the ML models see Table 6.2.

Feature subsets: To gain insight into the impact of each feature category on the classification performance, we created subsets by adding one or more feature categories to the previous subset, resulting in eight sets of features that we used to train the ML models, starting with the initial subset containing only the basic feature categories and ending with the final set containing all extracted features.

Table 6.3: Number of classes for each classification problem, including train, validation, and test dataset size.

Classification problem	Number of classes	Number of instances		
		train	validation	test
Counselor behavior				
Fine-grained predictions	17	7,341	918	918
Evocative vs. non-evocative	2	7,341	918	918
MI-congruent vs. MI-incongruent	2	9,700	1,212	1,213
Client behavior				
Fine-grained predictions	4	9,485	1,186	1,186

Train-validation-test split

For each classification problem, we split each group of chat messages - stratified by class distribution - 80/10/10 percent to create training, validation, and test datasets. To rightly measure model performance, it is essential to hold out data. We used the training set exclusively to train the models, evaluated the training progress on the validation set, and obtained the final performance using the test set. Table 6.3 shows the number of classes and instances for each classification problem.

Learning algorithms

We trained and evaluated four different ML models and one deep learning model for each classification problem: a Random Forest (RF), a SVM, a k-nearest Neighbors (kNN), a Decision Tree (DT), and a *pre-trained transformer model*.

We chose a pre-trained transformer model to overcome the limitations of regular deep-learning model architectures. Transformer models are a type of deep learning network that can be pre-trained on a large amount of data and then fine-tuned on a smaller, more specific dataset to make predictions. By pre-training on a large dataset, the model can learn to understand the structure and patterns of language, making it easier to adapt to new domains, which enables the training of complex models with limited data. Researchers showed that the BERT model [98] suits this approach particularly. We used a variant of BERT, called BERTje, which already has been pre-trained on a large Dutch text corpus [95], and fine-tuned BERTje on our domain-specific dataset for each classification problem.

A grid search technique was employed to select the best model parameters, as initial testing showed that the parameter values could severely impact model performance. The final analysis excluded compensating for the imbalance of the class labels in the data, as initial testing also showed that it did not lead to differences in the results. To account for this, we evaluated the models using statistics that can take class imbalances into account.

We used five-fold cross-validation to validate the models and applied min-max scaling before training. We implemented all ML models in *Python 3.8* and implemented the fine-tuning of BERTje using *PyTorch Lightning*.

Evaluation metrics

Computing the *confusion matrix* and conducting an *AUC-ROC analysis* allowed us to assess the classifiers and obtain visual and statistical insights into their predictive performance. We also quantified the *kappa statistic* and *accuracy* for the best-performing models. We extracted the probability distribution of the predictions from all classifiers to compute the *sample average F1 score*. The probability distribution indicates the confidence or likelihood of a specific model prediction.

Baseline: The baseline score provides a required point of comparison when evaluating all predictive algorithms for a classification task. We consider predicting the majority class as a baseline, meaning that we select the prediction class with the most observations and use it as the outcome for all predictions. We expect the predictive models that learn from the data to perform substantially better.

Validity: We used identical statistics, training, validation, and test samples to evaluate the trained models, making the validation of the results comparable across all models.

Explainability

To interpret the output of the models, we employed Shapley Additive Explanations (SHAP) [76] as a method. SHAP provides a way to obtain the contribution of each feature in the model's prediction for a particular input. Values provided by SHAP represent a feature's average marginal contribution towards the difference between the predicted output and the model's expected output. A higher value indicates a higher contribution to the output and interprets it as a more important feature.

6.3 Results

6.3.1 Algorithm performance

In this section, we present a comprehensive evaluation and performance analysis of the ML models and the transformer model BERTje, across all four classification tasks. We further interpret the model predictions by deploying SHAP and laying out the most occurring word combinations for each prediction class.

Classifying counselor behavior

Fine-grained predictions: Figure 6.1 presents a performance comparison of the learning algorithms using the best parameters for classifying counselor behavior. The reported scores represent the average of five repeated runs for each model. The SVM model ($\gamma = 0.1, C = 10$) showed the highest F1 score of 0.63 among all ML models. RF and SVM models outperformed the DT and k-NN.Incorporating textual information and word embedding features resulted in the highest increase in the performance of the ML algorithms. Among all models tested, the transformer model BERTje achieved the highest performance with an F1 score of 0.73. Table 6.4 shows a detailed model performance evaluation of BERTje. With an accuracy of 0.72, kappa statistic of 0.69, and AUC score of 0.95, it's results represent a 350% improvement in accuracy from the baseline. The AUC scores per class ranged from 0.89 to 0.99. The model performed best on *Fillers* and *Affirmations* with an AUC score of 0.99 and lowest on Advise without Permission (AWP) and

Table 6.4: BERTje: detailed model performance evaluation on all classification tasks.

	Evaluation metric						
Classification task	*F1*	*AUC*	*Sample avg. F1*	*Micro avg. AUC*	*Macro avg. AUC*	*Accuracy*	*Kappa*
Counselor behavior			0.73	0.96	0.95	0.72	0.69
Advise with Permission (AWP)	0.50	0.94					
Advise without Permission (ADW)	0.39	0.89					
Affirm (Aff)	0.86	0.99					
Closed Question	0.85	0.97					
Confront (Con)	0.31	0.89					
Emphasize Control (Econ)	0.38	0.91					
Filler (Fill)	0.80	0.99					
General Information (GI)	0.67	0.96					
Open Question (OQ+)	0.64	0.95					
Open Question (OQ−)	0.69	0.97					
Open Question (OQ0)	0.72	0.94					
Permission Seeking (Perm)	0.71	0.96					
Reflection (+)	0.51	0.93					
Reflection (0−)	0.76	0.97					
Self-Disclose (Sdis)	0.57	0.96					
Structure (Str)	0.84	0.97					
Support (Sup)	0.69	0.95					
MI congruency			0.88	0.94	0.92	0.87	0.65
MI-Congruent (X MI+)	0.91	0.92					
MI-Incongruent (X MI−)	0.76	0.92					
Evocative language			0.90	0.96	0.92	0.90	0.66
Evocative	0.73	0.92					
Non-evocative	0.94	0.92					
Client behavior			0.71	0.90	0.89	0.70	0.55
Ask	0.81	0.99					
Follow/Neutral (FN)	0.61	0.81					
Change Talk (X Csa+)	0.66	0.87					
Sustain Talk (X Csa−)	0.74	0.89					

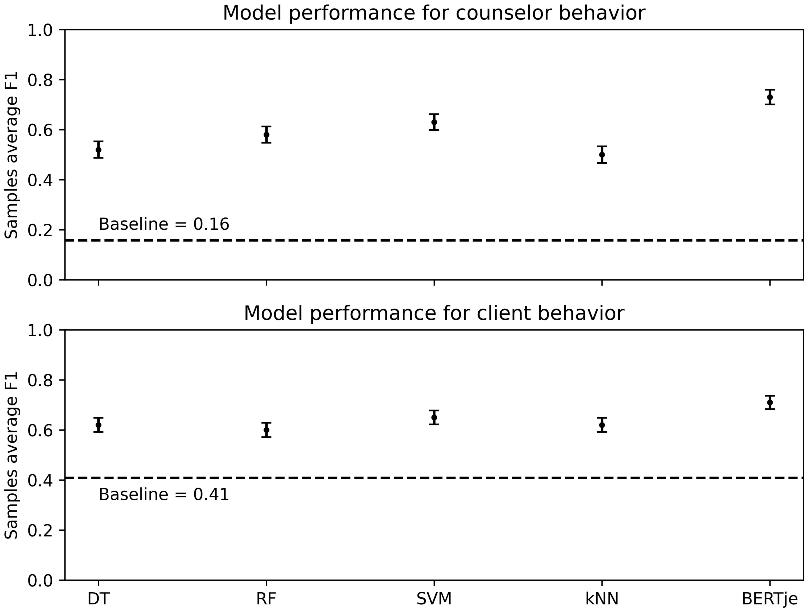


Figure 6.1: Learning algorithm performance for predicting client- and counselor behavior.

Confront (Con) with an AUC score of 0.89. Errors mainly occurred when predicting neutral, open questions as positive open questions (14 errors), positive reflections as neutral or negative reflections, and open questions as closed questions (17 errors).

MI congruency: The finetuned BERTje model achieved a sample average F1 score of 0.88 in accurately predicting counselor behavior as either MI-congruent or MI-incongruent (see Table 6.4). Additionally, it demonstrated a high accuracy of 0.87, accompanied by a kappa value of 0.65 and a macro average AUC score of 0.92. These results signify an accuracy improvement of 20.8% compared to the baseline performance (accuracy = 0.72).

Evocative language: The finetuned BERTje model achieved a sample average F1 score of 0.90 in accurately predicting whether counselor language is evocative or non-evocative (see Table 6.4). Moreover, it demonstrated an accuracy of 0.90, a kappa value of 0.65, and a macro average AUC score of 0.92. These results signify an accuracy improvement of 9.8% compared to the baseline performance (accuracy = 0.82).

Classifying client behavior

Figure 6.1 also shows a performance comparison of the learning algorithms using the best parameters for classifying client behavior. All models improved classification performance compared to the baseline. The best ML model was an SVM model ($\gamma = 0.1, C = 10$), reaching a sample average F1 score of 0.65. BERTje outperformed all ML algorithms, reaching a sample average F1 score of 0.71 with an accuracy of 0.70, Cohen's kappa = 0.55, and a macro average AUC score of 0.89. These results indicate an accuracy improvement of 70.7% compared to the baseline. The AUC scores per class range from 0.81 - 0.99 (see Table 6.4). Although the lowest occurrence across the client messages, BERTje predicted the code *Ask* best (AUC = 0.99). *Follow/Neutral* was the hardest to predict (AUC score = 0.81). We observed that errors mainly occurred in predicting *Follow/Neutral* messages as *commitment language*.

6.3.2 Feature contributions

According to the SHAP feature importance analysis conducted on the best-performing ML models, word embedding features held significant dominance. Furthermore, the number of question marks in a message emerged as a consistently influential factor for client- and counselor behaviors. Table 6.5 shows the features that contribute most to the predictions of each class individually. Moreover, this table shows the top word combinations reflecting the language character of different client- and counselor behaviors. The inferred prediction classes associated with the MI-SCOPE codes are generally interpretable. For example, client ambivalence becomes clear when counselors use reflections with word combinations such as *"on one side"*, *"on the other side"*, and *"conflicted"*. Concerning client commitment language, negative sentiment, and negations contributed to both Sustain- and Change Talk. When these features were *present*, client language was more likely to be associated with Sustain Talk rather than Change Talk. Contrarily, the *absence* of these features indicates more association with client Change Talk.

Table 6.5: Most influential features and word combinations contributing to the prediction outcomes and language character per class for counselor- and client behavior.

Class	Highest feature importance	Most occuring word combinations
Counselor behavior		
Advise with Permission (AWP)	# lowercase letters, # vowels	*seeking distraction; own environment; I think that; seeking contact; thoughts; express emotion, pleasant manner; creative; sports; general practitioner*
Advise without Permission (ADW)	# question marks	*I think that; how/what about; maybe it is good to; try to hold on; seek distraction; let it sink in; in any case; call 911 (Dutch: 112)*
Affirm (Aff)	positive sentiment, subjectivity score	*good for you; very wise of you; how great; seems like a good idea; good to hear*
Closed Question	# question marks	*did I get that right; do you ever; do you think that; do you also have; is this something to; are you still there; would you manage to; does your therapist know*
Confront (Con)	# question marks, neutral sentiment	*after hearing you; I think you; sounds like; I can imagine; a long time; crisis service; suicidal thoughts*
Emphasize Control (Econ)	use of pronouns	*what would you like to discuss; what do you need the most; look together; a friendly and listening ear; is there still something else*
Filler (Fill)	# stopwords, sentence length	*welcome to the chat; thank you for waiting; thank you for your openness; you're welcome; no problem; you too; okay; hmm*
General Information (GI)	use of punctuation, # special characters	*online therapy; regular psychologist; website; via email; five working days; finding information; registration; https://www.113.nl*
Open Question (OQ+)	# question marks, positive sentiment	*what would you need; what do you like to do; what could it bring you; what do you think of …; how would you; what do you usually do*
Open Question (OQ−)	# question marks, negative sentiment	*what happened; how come; what makes you think that; what's going on; what is the worst that could happen; what can you tell more about …*
Open Question (OQ0)	# question marks, use of adjectives	*how does this feel for you; what is your point of view about; how would it be like to …; what do you think; what makes that; how would you*
Permission Seeking (Perm)	use of the word "I", # unique words	*shall we discuss our ideas together; is it okay for you if; are you comfortable with this; is it an idea to; share information*

Continued on next page

Table 6.5 – continued from previous page

Class	Highest feature importance	Most occuring word combinations
Reflection (+)	use of the word "you", positive sentiment	*sounds like; you indicated that; you're describing; you feel; if I understand correctly; on one side; on the other side; conflicted; listening ear; look together; for now you want*
Reflection (0–)	use of the word "I", negative sentiment	*you feel drained; clearly, there's a lot going on; you've had some negative encounters; gone through a bad time; it feels like; suffering from suicidal thoughts; tension; restlessness*
Self Disclose (Sdis)	use of the word "I"	*from my own experience; I know; I see that you; I think; I hope you; for me; I am; I find it; oh sorry*
Structure (Str)	# question marks	*hi, you are speaking with . . . ; just a moment; I'll be right back to you; close the chat; read back our conversation*
Support (Sup)	neutral sentiment	*sorry to hear; sad to hear this; I understand your thoughts; that does sound like; I can imagine; good luck; get well soon*
Client behavior		
Ask (Ask)	# question marks	*what do you mean by that; what can I do; what should I; but how can I; what if; do you agree with; what kind of help*
Change Talk (X Csa+)	negative sentiment, negations	*good idea; very nice; I could try that; I think so; will help; talk about it; look for a distraction; listen to music; watch TV*
Follow/Neutral (FN)	# short words	*that's fine; I don't know if; nothing to worry about; I know; thanks for your time / help; yes; no; thanks for the conversation*
Sustain Talk (X Csa–)	negative sentiment, negations	*I don't want to; I'm afraid; when I'm not here anymore; I don't know how; I find it difficult; I feel really bad*

6.4 Discussion

6.4.1 Interpretation of the results

The results of this study demonstrate the potential of AI models, particularly the transformer model BERTje, in classifying MI behavior in online mental health helplines. BERTje outperformed all ML models tested, achieving high levels of accuracy across all classification tasks. Although ML models obtained lower performance than the BERTje model, their high explainability adds value for gaining a deeper understanding of language use concerning specific MI behaviors.

The successful application of a finetuned transformer model in classifying MI behavior is consistent with other recent studies, such as [164] and [148], who also used a finetuned transformer model to classify MI behavior in counseling sessions. Both studies also used some form of model interpretation to understand how the models make predictions and what features or words characterize each class. Our study extends this line of research by using a different dataset, coding scheme, and transformer model than the previous studies. Both related studies used the MISC codebook for data annotation and did not provide any estimates of coding reliability. In contrast, our study offers an in-depth account of the procedures and methodology, including reporting on coding reliability and fidelity measures. Studies that also utilized the MI-SCOPE (e.g., [101]) used a small dataset and obtained lower F1 scores than the current study, highlighting the importance of using larger datasets to improve the performance of AI models in predicting MI behavior.

Our study contributes to the growing evidence base for MI as an effective intervention for various health-related behaviors. We showed that the AI models can accurately identify the effective ingredients of MI, such as client change- and sustain talk, counselor affirmations, and reflection types - facilitating valuable counselor feedback. Furthermore, this study is the first to apply such a model to the domain of suicide prevention, which poses specific challenges and opportunities for MI. For example, counselors in a suicide prevention helpline need to adhere to MI but also balance building rapport, exploring ambivalence, focusing on engagement, collaboration, and empathy, and ensuring client safety.

6.4.2 Strengths and limitations

This study held several notable strengths. In our methodology, we adhered to the best AI practices. We used a hold-out test set to evaluate the performance of the AI models, providing a realistic estimate of their generalization ability. Using diverse statistics to evaluate the model performances, we make the validation process comparable across all models. We clearly described the analytic strategy, ensuring transparency and reproducibility of the research process substantiated by the comprehensive supplemental material.

In the context of generalizability, classifying MI behavior could relatively easily be deployed in other domains and other languages. These days, large language models are being pre-trained on many texts in multiple languages. From an AI point of view, implementing these methods in other online mental health helplines is relatively effortless.

While this study holds several strengths, there are also limitations. The dataset utilized in this study is relatively small, which could lead to higher variance in the test set. Some MI code groups were under-represented in the dataset (e.g., less evocative statements occurred in the text than non-evocative statements). This limited dataset size may restrict the model's ability to generalize to a broader range of MI conversations. It is important to note that the model's performance in this specific domain of suicide prevention does not guarantee its effectiveness in other settings. To assess and train the model performances in another domain's context, it is still necessary to gather domain-specific data.

A final limitation is the disregard for demographic traits of clients and counselors, such as age and cultural background. These characteristics could influence language use and model predictions. Further research is needed to refine the AI models with these factors, but these are not without ethical concerns [167]. Additionally, years of counselor experience and MI proficiency could affect model effectiveness, with less experienced counselors likely benefiting more from the model.

6.4.3 Implications for clinical practice

Leveraging AI models for clinical support

There are several potential ways to incorporate AI models into clinical practice to enhance MI proficiency in online helplines for mental health:

By integrating AI models into chat-based counseling platforms, counselors can receive instant feedback on their MI behavior during sessions. This feedback allows counselors to review their generated messages before sending them and make necessary adjustments, such as changing a closed question to an open one. Initial results suggest that counselors find such systems acceptable [144], but more studies are needed to evaluate the reception and impact of these tools in different settings and populations.

By offering post-session feedback and training to counselors, they can reflect on their performance and pinpoint areas where they may require additional training or support. By analyzing data from multiple counseling sessions, AI models can detect patterns or trends in counselor MI behavior and develop tailored training programs that offer recommendations for training or support. This integration can assist counselors in identifying areas where they may need to modify their approach or apply MI techniques more effectively.

Figure 6.2 shows a schematic overview illustrating the proof of concept of the support tool. Studies can investigate the feasibility of such a tool using a Wizard-of-Oz approach, where an experienced counselor acts as the support tool to simulate a best-case scenario. This setup could serve as a preliminary test for the viability of the support tool without requiring the development of a fully equipped AI tool for examining the potential advantages or disadvantages.

Madeira et al. [122] proposed such a tool, and Salmi et al. [144] examined it in focus groups with counselors and tested viability in a simulated environment. Studies emphasize the significance of tools that can help ease the workload of counseling [174, 178]. Several studies also highlight relevant elements concerning the viability of AI in mental health counseling [171, 174, 182]. Therefore, a comprehensive evaluation of feasibility is necessary.

Helpline administrators or supervisors could also use AI models to monitor and evaluate the quality of counseling services provided

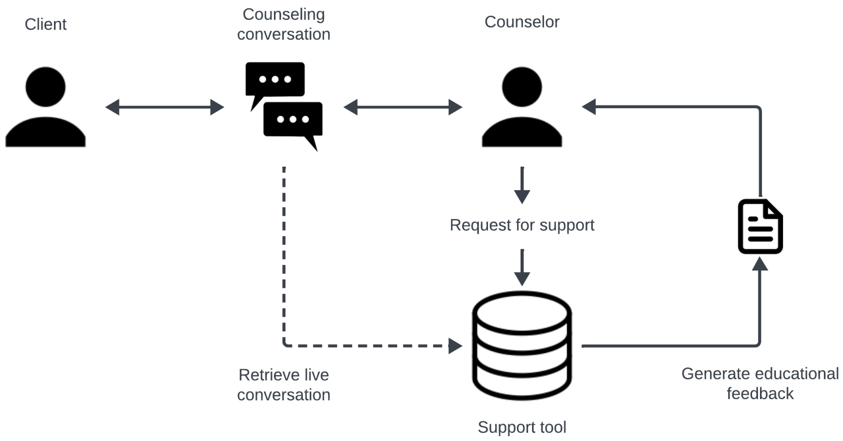


Figure 6.2: Schematic overview illustrating the proof of concept of the support tool.

by their organization. Using AI models to classify MI behavior and providing counselors with feedback offer a scalable and cost-effective solution for enhancing MI proficiency in helplines and other counseling settings. Many helplines struggle to find good staff, and the turnover is high, so reducing the *Time to Proficiency* is very valuable.

Evaluating the effectiveness of aI models in clinical practice

One way to evaluate the effectiveness of AI models in clinical practice would be to conduct pilot studies or randomized controlled trials analyzing changes in counseling outcomes over time. By comparing data from counseling sessions before and after the introduction of AI models, it would be possible to determine whether using these models leads to improved health outcomes for individuals seeking help.

Another way to evaluate the effectiveness of AI models would be to monitor changes in counselor MI behavior over time. By analyzing data from multiple counseling sessions, it would be possible to determine whether counselors who receive feedback from AI models improve their proficiency in applying MI techniques.

Feedback from counselors and clients could also provide valuable insights into the effectiveness of AI models in enhancing MI proficiency in chat-based counseling sessions. Counselors could provide feedback on the usefulness and accuracy of the provided feedback by the AI models, while clients could provide feedback on the quality of their interactions with counselors.

Next steps to take

Clinicians or researchers interested in leveraging AI for their specific use case in online helplines for mental health can already take initial steps to get started. A crucial first step in leveraging AI to enhance MI proficiency is to collect data from counseling sessions, such as chat transcripts and relevant metadata. Another important step is connecting with other clinicians and researchers in the field. By joining an active community of professionals working with MI and AI, one can benefit from the knowledge and resources created by others. These resources may encompass pre-trained AI models, guidelines for collecting and analyzing data, and opportunities to collaborate and share knowledge with others in the field.

6.4.4 Future directions

While the results are promising, additional research is needed to evaluate the performance of these models on larger datasets with sufficient representation of each class. Future studies may explore alternative modeling techniques that better capture the conversational structure in classifying MI behavior. For instance, graph-based models can store information about the relationships between messages within and across conversations.

While this work and earlier research successfully quantified and validated the technical aspects of MI, it is also relevant to consider fundamental principles of MI. These principles contain the conversational processes that guide interactions between counselors and clients. Adhering to these processes ensures that counselors do not exceed the level a client is comfortable with and adapt their behavior appropriately based on the specific context. For instance, an appropriate question during the *engaging* process (building rapport between client and counselor) may become counterproductive during the *evoking* process, where the main goal is to elicit change talk. In these processes, concepts like collaboration, engagement, and empathy also play a significant role. Recent work is already exploring integrating these concepts into AI models in the context of online helplines for mental health [130, 178, 183].

Furthermore, applying the methods used in this study to other languages and institutions could provide valuable additional validation of the study findings. A final point of future work would be to investigate and measure the potential improvement of MI quality in the chat helpline before and after counselors used MI insights during their conversations over time. This could also be combined with a support tool for MI feedback, in a randomized control trial.

6.5 Conclusion

The results of this study demonstrate that AI techniques can accurately classify MI behavior, indicating their potential as a valuable tool for enhancing MI proficiency in online helplines for mental health. Provided that the dataset size is sufficiently large with enough training samples for each behavioral code, these methods can be trained and applied to other domains and languages, offering a scalable and cost-effective way to evaluate MI adherence, speed up behavioral coding, and provide therapists with personalized, quick, and objective feedback.

Chapter 7

Discovering effective counselor utterances with interpretable deep learning

Salmi, S., Mérelle, S., Gilissen, R., van der Mei, R. D. & Bhulai, S.

Under review.

Abstract

Background
To provide optimal care in a suicide prevention helpline, it is important to know what contributes to positive or negative effects on help seekers. Helplines can often be contacted through chat services, which produce large amounts of text data, to use in large-scale analysis.

Methods
From August 2021 until January 2023, help seekers (*N* = 6903) scored themselves on factors known to be associated with suicidality (like hopelessness, feeling entrapped, will to live, etc) before and after a chat conversation of 113. ML text analysis was used to predict help seeker scores on these factors. The model was interpreted, to show which messages of the helpers in a conversation contributed to the prediction.

Results
According to the ML model, positive affirmations and expressing involvement of helpers contributed to improved scores of help seekers. Use of macros and ending the conversation prematurely, due to the help seeker being in an unsafe situation, had negative effects on help seekers.

Conclusion
This study reveals insights for improving helpline conversations, emphasizing the value of an evocative style with questions, positive affirmations, and practical advice. It also underscores the potential of machine learning in helpline analysis.

7.1 Introduction

An important question that has yet to be answered regarding helplines is what counselling approach is effective to take. Helplines are often anonymous however, which makes it difficult to do evidence-based research, and little is still known. Several studies have been conducted on the Crisis Text Line to identify the characteristics of help seekers and their perception of the helpline's effectiveness [155][163]. Furthermore, Gould et al. [53] examined call reports of help seekers calling helplines in the National Suicide Prevention Lifeline network. In studies by Mokkenstorm et al. [63] and Mishara et al. [24] helpline chat logs were annotated and analysed for the purpose of gathering empirical evidence. A downside of these approaches is that manual annotation of chat logs is often time-consuming work and not a lot of available data is left unused.

113 uses pre- and post-conversation questionnaires to assess the help seeker's mental well-being [158], i.e. questions related to associated suicide risk factors, such as hopelessness, entrapment, perceived burdensomeness and thwarted belongingness. This provides a method to gauge a conversation's outcome. By training a classification model to predict chat outcomes based on the content of the chat conversation, insights can be gained from looking at the model's functioning. When help seekers have lower scores on the questionnaire after the conversation, this would indicate they are less distressed, and the conversation would be classified as positive and vice versa. A better understanding of what contributes to a positive conversation can help inform helplines and possibly result in actionable recommendations for helpline policy.

However, training a classification model on this data is not a trivial task. Not unlike sentiment analysis of large documents, a decent level of accuracy is difficult to achieve. Furthermore, the model should be interpretable, such that insights can be gained from the relation of the text content to the classification, i.e., which parts of the conversation have more impact on the model's output.

The length of the training data that is used by the best models is in the order of a couple of sentences. This is a main limitation of the transformer architecture, because computing space and time scales quadratically with sequence length. This means that text must be truncated before it can be interpreted by the model, and truncation means information will be lost. The long length of conversations in a crisis line easily leads to a lot of loss of information.

Hierarchical models have been a method to get around this limitation. By first applying the model to a subsection of the sequence, a representation for that subset can be learned. A chat conversation can quite easily be segmented into individual messages or groups of concurrent messages, which becomes the subset that we can learn a representation for. By keeping the second level in the hierarchy simple, this approach also allows to more easily create an interpretable network.

In this research, we trained and compared several hierarchical models that leveraged pre-trained BERT models, with the goal of gaining insights into the quality of helpline conversations. Among these models, we included a weighted average model with conversation participant masking. The results showed that this model performed comparably to the best hierarchical models while adding additional interpretability. Using the weights provided by this model, we rank important messages of a test set, which contributed to improved and not improved scores after helpline conversations.

7.1.1 Related work

Many state-of-the-art language models that have been developed in recent years rely on transformers [180]. In the original paper, the transformer was developed as an encoder-decoder network for the task of language translation. Devlin et al. [98] adapted the encoder section of the transformer to create high-quality embeddings. Dubbed BERT, this network, or variants thereof, were applied to get state-of-the-art performance on many NLP tasks [105]. Many pre-trained variants of these networks have been made available. For Dutch language tasks there are two variants trained: Bertje, based on the BERT network and RobBERT, based on the RoBERTa network.

Due to the nature of the attention mechanism, a straightforward assumption to make is that the attention weights directly relate to token importance. However, this assumption has been frequently questioned. Serrano and Smith [110] found that attention weights only noisily predict importance. Jain and Wallace [103] argue attention weights do not provide explanations, while Wiegreffe and Pinter [112] in turn challenged their claims. They argue that there is a time and a place for it, and provided tests to determine when attention can be used in such a way. A frequently suggested alternative to attention weight as an importance metric is Gradient-based saliency [111] [115]. However, even saliency maps have limitations [115]. Local Interpretable Model-Agnostic Explanations (LIME) [66] is a general method for explainability that can also be applied to NLP. LIME generates explanations for complex models by locally approximating their behavior with simpler models.

A main limitation of transformers are very long-range dependencies, because self-attention scales quadratically in the length of the sequence. This $O(n^2)$ time and memory complexity means that models often limit or truncate the input to a certain length. Issues arise with a corpus of documents that almost always exceed this limit, such as the one used in this study. Therefore, several adaptations of the transformer method have been proposed to deal with this issue. Longformer [116], got around this limitation by using windowed attention, combined with a limited number of global task-specific tokens. An alternate approach is used with hierarchical networks [71]. By first computing a fixed representation for a smaller section of the sequence, these representations are the used as input for another sequence based approach. In this method, a sequence has to be split up in some way. Often times, paragraphs are used as the delimiter, however in the case of conversations a message or utterance could also be appropriate.

In the domain of text analysis for healthcare, several applications of transformers have been used to gain insights into healthcare text data. Gao et al. [100] found that pre-trained BERT models did not outperform simpler methods for medical document classification. These simpler methods consisted of a convolutional neural network and a hierarchical self-attention network, which had similar performance while having fewer learnable parameters. Ilias and Askounis [157] used LIME to find influential words of BERT classifications of Dementia transcripts.

7.2 Methods

7.2.1 Task definition

We modelled the problem of predicting the outcome of a helpline chat conversation as a binary classification task. We compared the scores of the questionnaire before and after the conversation. The classification outcome was defined as whether the help seeker's score on the questionnaire for suicide risk factors improved or did not improve.

7.2.2 Data

The data consisted of chat conversations of a suicide prevention helpline. Between April 1st 2021 and March 31st 2022, help seekers of this help-line were asked to fill in a short questionnaire on suicide risk factors before, and after the conversation with a counselor. Conversations that already started at the best possible value for the questionnaire before having the conversation were left out of the dataset. Conversations in the suicide prevention helpline also included a triage, where the help seeker was screened for safety. The triage part of the conversation was left out of the dataset as well. Without the triage, conversations had 64 messages on average. Due to a large class imbalance between improved and not-improved pre-post scores for conversations, we rebalanced the data. Randomly, samples from the larger positive class were removed, so that it matched the size of the negative class. The resulting final dataset used 6,000 chat conversations for training and 903 conversations as a test set.

7.2.3 Chat message embedding

The individual messages were embedded using a pre-trained RoBERTa network called RobBERT [118]. This network was subsequently fine-tuned on the chat conversations using a triplet-loss strategy. The models that are described in the remainder of this section used this network to embed individual chat messages first. A message embedding was created using a pooling layer, resulting in a matrix $C \in \mathbb{R}^{b \times l \times d}$, where l is the length of the sequence, d the embedding size of the pre-trained network, and b the batch size.

7.2.4 Weighted average

To improve explainability, we used a simpler adaptation of the attention mechanism. The weighted average is defined as (7.1).

$$\textbf{Weighted Average}(C) = \text{softmax}((CW_k^T + b_k)^T)(CW_v^T + b_v) \quad (7.1)$$

Here $C \in \mathbb{R}^{n \times d}$ is the matrix of input embeddings with embedding dimension d, from messages belonging to a conversation of length n. $W_k \in \mathbb{R}^{1 \times d}$ and $W_v \in \mathbb{R}^{d \times d}$ are learnable weight matrices. This approach can also be described as simplified version of dot product attention, where only a single class token attends to the sequence. This removes the need for the projections Q and K. This weighted averages results in a d dimensional vector which is used as input for a final feed-forward layer for classification. Because we were also interested in the speech of the counselor in particular, one additional adaptation we made is the inclusion of participant masking. Each weighted average is conditional on the sender. So in a conversation, each weighted average only considered the messages of each participant. This was done by using multiple weighted averages and masking the logits of the weights for the weighted average which corresponded each participant. As is common in transformer models, we also used multiple heads, which meant the model created multiple weighted averages. The final heads were then concatenated and projected to a classification output.

Before the message embeddings were combined into the weighted average, the weights were first masked. We created two masks, one for only the counselor message and one for only the help seeker messages. This resulted in the weighted average only being an average of either the counselor or help seeker. The counselor and help seeker each had the same number of heads.

7.2.5 Other hierarchical models

We also applied the same hierarchical method, of embedding the chat messages and hierarchically classifying these shorter inputs with three other methods. We applied a four layer LSTM [8] on message embeddings. We also applied four transformer embedding layers. A trainable class vector was concatenated to each sequence, which was pooled as the output. The final method applied a simple average of all message embeddings over the sequence dimension. The outputs of these models were fed in the same feedforward layer as the weighted average method.

7.2.6 Baseline models

We applied several additional preprocessing steps for the baseline models. All words were also lowercased, lemmatized and all special characters and punctuation were removed as well as stop words. During tokenization we limited the number of tokens 2,000. We vectorized the chat conversations using TF-IDF [36]. Finally, each embedded conversation was trained on a support vector machine [25].

Furthermore, the dutch BERT model, RobBERT, was used as another baseline model. Because it has a maximum length of 512 tokens for the text input, the chats were truncated the maximum length. Two RobBERT models were fine-tuned, one where the start of the conversation was truncated, and one where the end was truncated.

7.2.7 Explainability

To gain insights into the workings of the network, we employed two techniques. First we used the weights of the weighted average model. The assumption was that messages with a higher weight were of higher importance to the final decision and, therefore, more

Table 7.1: Model performance on the test set of the suicide chat classification task

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.638	0.635	0.632	0.633
BERT truncated end	0.570	0.556	0.699	0.620
BERT truncated start	0.629	0.605	0.743	0.667
Hierarchical Average	0.640	0.621	0.721	0.667
Hierarchical Weighted Average	**0.683**	**0.679**	0.697	**0.688**
Hierarchical LSTM	0.672	0.674	0.668	0.671
Hierarchical Transformer	0.638	0.612	**0.754**	0.676

important to the result of the conversation. As a second technique, we applied LIME [66] to the models. For this approach, we left out counselor messages one at a time to compute the difference in loss. A larger difference indicates more importance to the classification.

7.2.8 Ethics approval

The study protocol is performed in accordance with the relevant guidelines. This study was reviewed and approved by the Medical Research Ethics Committee of Amsterdam Universitair Medische Centra (registration number: 2021.0447).

7.3 Results

7.3.1 Model performance

Table 7.1 shows the performance scores on a held out test set. The Hierarchical Weighted Average model was the best performing model with an accuracy of 0.683 and also the highest F1 Score of 0.688. This was closely followed by the Hierarchical LSTM model, which had an accuracy of 0.672. The results for the Hierarchical Transformer model and Hierarchical Average did not perform as well, with accuracies of 0.638 and 0.640.

The SVM model had an accuracy of 0.638, which was the lower than the hierarchical models. The two truncated versions of the BERT model had accuracies of 0.570 and 0.629 for the truncated end and truncated start models, respectively. This suggests that the information in the truncated text was most likely insufficient compared to the hierarchical models that do not have the ability to attend to words from different messages, but overall have more information available.

Overall, our results suggest that the Hierarchical LSTM and Hierarchical Weighted Average, outperformed other models for the task classifying suicide helpline chat conversations.

7.3.2 Model explanations

The Weighted Average model was the overall best performer in terms of accuracy and F1 Score. Because it was more interpretable than the Hierarchical LSTM and the BERT networks, it was the obvious choice to extract explanations from. The explanations were compiled from a test set only, and a sub-selection of the data was made, with only the correctly classified samples, and where the model was confident in its output. This confidence was measured through the logit output of the model. A logit value close to 1 corresponds to the classification of a chat conversation that resulted in an improved score, and closer to 0 for the opposite case. Figure 7.1 shows a histogram of the logit outputs of the model for this test dataset. Two peaks can be seen for the correctly classified samples. This shows that, while there are still chats that are difficult to categorise, there is a clear set of chats where the model is confident for either class. We can also see that the model was slightly more confident for not-improved chat conversation than for improved ones. We chose values below 0.2 and above 0.8 as subsets to extract the explanations for.

Using the weights from the hierarchical weighted average model, we compiled the most influential message from counselors. The messages selected as influential were messages with a weight one standard deviation above the mean. The weights used for this purpose were from the heads that were masked for the help seeker and thus only contained non-zero values for counselor messages. Furthermore, LIME was also used in combination with the Weighted Average model to obtain explanations. This section describes the

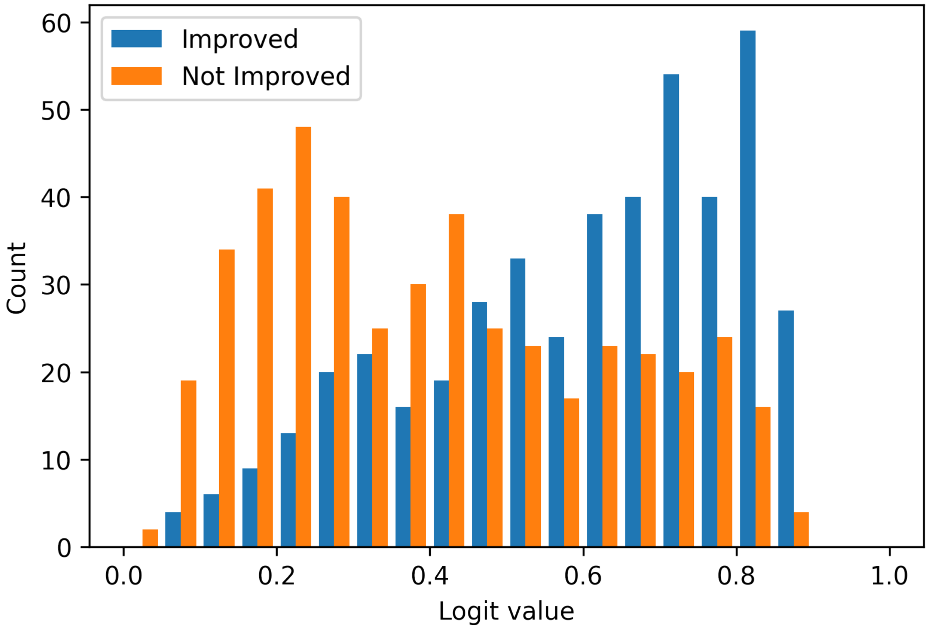


Figure 7.1: Histogram of logits for the Hierarchical Weighted model

outcomes based on observations by the authors and two senior psychologists from the helpline.

Conversations without an improved suicide risk score

For conversations that did not improve, we identified three distinct situations that emerged from the influential messages. The first and most common situation was when a conversation ended prematurely. In such cases, the counselor would typically try to redirect the help seeker to alternative channels for assistance, such as a general practitioner, a different helpline, or emergency services. Alternatively, in some cases the counselor suggested to contact the helpline at another time or to apply for an online therapy service. The second situation involved messages where the counselor was unable to respond promptly to the help seeker due to a high volume of ongoing conversations. The counselor apologized to the help seeker for the delay and sometimes mentioned that the helpline was particularly busy, and the counselor was dealing with multiple help seekers simultaneously. The third situation included the counselor not connecting properly with the help seeker. This was expressed in the use of macros and lists. The macros would often include a standardized set of options for the help seeker to consider or sometimes a set of websites and resources to visit. Sometimes this was also expressed as the counselor not properly listening to the help seeker.

Conversations with an improved suicide risk score

The results showed that conversations with improved scores had a wider range of responses compared to those that did not improve. However, we identified two frequently recurring situations in conversations that showed an improved score. In the first situation, the counselor provided positive reinforcement to the help seeker. During these conversations, the counselor would typically use supportive language, such as showing empathy, offering praise, and expressing happiness for the help seeker. In the second situation, the counselor expressed involvement. For example, the counselor would think along with the help seeker, and provide concrete solutions to the help seeker. These solutions could include specific actions or resources tailored to the help seeker's individual situation. The counselor would provide the help seeker with practical steps that could be taken, or resources specific to the help seeker's situations. Lastly, two less often recurring situations included situations where the counselor would ask open ended questions, as well as show respect for the autonomy of the help seeker by asking what they wanted to do.

7.4 Discussion

This study compared the performance of different models for classifying suicide helpline chat conversations and found that the Hierarchical Weighted Average model had the best performance. This study also extracted explanations from the Hierarchical Weighted Average model and identified three distinct situations for conversations that did not improve and two clear recurring situations for conversations that showed an improved score.

The results showed that the model had an easier time determining when a conversation would not lead to an improvement in the risk factors. This was also apparent in the explanations where clear and easy distinctions in the output could be made, whereas this was not as easy to do in the case of positive examples.

Research by Mishara et al. [24] found that collaborative problem-solving significantly predicted positive outcomes in helpline calls. In line with these findings, our study showed that messages with positive reinforcement and concrete solutions contributed to positive outcomes in chat conversations. Furthermore, Côte and Mishara [152] found through qualitative analysis that reinforcing a strength or a positive action a significant predictor was for increased scores on a pre-post questionnaire in a text-message helpline setting. This is in line with our finding that positive reinforcement was a frequently occurring impactful message. In a qualitative study by Gilat and Rosenau [33] analyzed volunteers' perspective of effective methods in helpline conversations. Among their findings they identified practical advice as an effective strategy. Building rapport was another aspect of note that their study identified. Because building rapport is highly specific to the individual this might be something our method was not able to generalize and pick up on. However, positive reinforcement could also have been a contributor to building rapport.

Overall, these findings highlight the potential of using machine learning models to analyze suicide helpline chat conversations and provide insights into the most influential messages. This allows helplines to be more informed and possibly enable them to improve helpline quality.

7.4.1 Limitations

While this study sheds light on influential messages in suicide chat conversations, there are three key limitations to be considered. First, there are general limitations of machine learning. The classification task was found to be difficult, as indicated by the 68 percent accuracy rate that was achieved. This suggests that the current models have room for improvement. There might still be relationships that the current models were not able to capture. It could also be that there is considerable noise in the dataset as a consequence of the self-reporting of the outcome measures by help seekers, which might not have been reliable for every help seeker. Furthermore, the indicated influential messages might be messages that are a result of a different action. For example, we saw multiple situations where a message was indicated as influential where the counselor expressed gratitude for a compliment. This was most likely the result of the help seeker being grateful for something, however it does not necessarily mention what the help seeker was grateful for. Second, a limitation of this study is the challenge posed by modeling a large amount of text. Current methods have either limitations in capturing dependencies over long ranges or in exceeding maximum memory thresholds, which was the case with the chat conversations used in our dataset. Therefore, hierarchical models were used, which had the limitation that dependencies between words from different messages were not captured. Third, a limitation of using chat messages as output to determine categories of influential messages is the need for human judgement. This introduces subjectivity and the potential for bias, as different judges may interpret the same messages differently, or possibly miss a connection between the different messages.

7.4.2 Future work

Considering the findings presented in this study, we identified three potential directions for future work that could further enhance the classification and identification of influential messages in suicide helpline chat conversations. First, while larger models have the potential to improve performance, explainability needs to be considered as well. The use of larger models can sometimes lead to decreased interpretability, and it may be challenging to identify the most influential features that contribute to the classification of a message. Therefore, future research could explore the use of models such as Longformer, which are designed to handle long sequences of text through windowed attention and global attention for the class.

This global attention can possibly be leveraged for explainability. Second, with additional computational resources, another potential area of research is to forgo the use of sentence embedders, and input messages directly into a transformer model. This approach could potentially improve the performance of the model by better capturing individual sentences, rather than relying on message embeddings that are not trained for the specific task. Third, in addition to model improvements, future research could explore additional processing techniques of influential messages, such as clustering. Clustering could be used to group similar messages together, allowing for an analysis of influential messages. This could be useful for easier identification of patterns in influential messages.

7.4.3 Practical implications

Engaging with help seekers expressing suicidal thoughts while recognizing they can be better helped elsewhere is important. However, counselors should be mindful of empathetically guiding them toward the appropriate channels. It is important to keep validating their emotions and ensuring they feel supported rather than dismissed. Standardized responses from macros can be beneficial in the right circumstances. If used without having good rapport with the help seeker they can appear distant. Being transparent with the help seeker about the use of macros is important, as well ensuring good enough rapport has been established with personal responses before using standardized responses. Collaborative problem solving and building rapport are proven ways to foster better conversations. Positive reinforcement might be another method that counselors can employ. Including positive reinforcement more regularly in their responses might be beneficial to helpline conversations.

7.5 Conclusion

This study compared the performance of different models for classifying suicide helpline chat conversations and found that a Weighted Average model using message embeddings performed the best. This study is unique compared to other studies that aim to gain insight into the quality and effectiveness of suicide prevention helplines. Many studies use questionnaires to evaluate implemented counseling approaches. In this study, we identified influential messages that contributed to better or worse scores on a suicide risk questionnaire, through a machine learning approach. This initial application showed that we could extract explanations from the model and identified distinct situations for improvement and deterioration of help seekers' emotional states.

Chapter 8

Discussion

8.1 Main findings of the dissertation

The work in this dissertation highlights the key discoveries and implications of the research, providing a comprehensive overview of how NLP techniques can enhance suicide prevention helplines.

Initial evaluation in Chapters 2 and 3 has shown important strenghts and weaknesses of using NLP systems to aid counselors by providing them with conversation suggestions. We observed that support tools powered by deep learning transformer models can provide related content based recommendations. However, a clear effect on the self efficacy of counselors was not found. Longitudinal studies are required to fully understand the effectiveness of these real-time NLP systems. Future research should focus on conducting additional controlled trials and gathering qualitative and quantitative data from a wide range of helpline scenarios. This will help in identifying the specific conditions under which NLP tools are most effective and the potential areas for improvement, ensuring that the technology meets the nuanced needs of helpline operations.

Chapter 4 presented a comparison of the best methods when using topic modeling for helpline conversation data. Chapter 5 applied the findings and examined the shifts in conversation topics during the COVID-19 pandemic, highlighting increased discussions on isolation, health anxiety, and financial stress. This temporal analysis enables the possiblity for helplines to adapt training and support strategies in response to evolving needs, making flexible policies and continuous monitoring to stay ahead of emerging mental health trends a possibility.

Chapter 6 explored the ability to automatically detect and categorize the use of specific MI techniques and has shown this has become a feasible pursuit. ML can potentially provide valuable support to counselors. Enhancing these models will support the ongoing professional development of counselors and ensure a high standard of care for help seekers. The utilization of ML classification methods to analyze conversation data and provide insights into interactions between counselors and helpseekers that should be sought after, or avoid, has been highlighted as a potential area for significant impact in Chapter 7. By advancing these classification methods, more insights can be gained, which can potentially lead to on-the-ground measures being inplemented top-down in the helpline. Furthermore, systems may be developed that assess the quality of conversations in real time and provide counselors with actionable feedback. This targeted feedback can help counselors become aware of certain strategies during conversations.

8.2 Research Questions

8.2.1 How can NLP techniques support suicide prevention helplines?

This question was explored throughout the dissertation but the most concrete example can be found in Chapters 2 and 3, where real-time recommendations during chat interactions were studied as a method to combat the counselor's cognitive load and maintain helpline quality. Leveraging transformer-based models like BERT, the system can suggest relevant past conversation segments, aiding counselors in more difficult situations. On the administrative side, topic modeling provides valuable insights into the operational aspects of helplines. By analyzing conversation through this method, one can identify common themes and issues.

8.2.2 What new insights can be gained from helpline data with NLP techniques?

NLP techniques provide valuable insights from helpline data, particularly in understanding behavioral and emotional trends. Temporal analysis allows for examining changes in the frequency and nature of specific topics over time, offering insights into how external events, like the pandemic, impact help seekers. NLP techniques also enhance the understanding of help seeker needs. They can identify trends and patterns in help seeker behavior and concerns, informing the development of targeted interventions and support strategies. By continuously analyzing incoming data, NLP systems help helplines adapt to changing needs and crises, such as during the COVID-19 pandemic when issues like anxiety and isolation became more prevalent. Furthermore, NLP can help detect and provide insights on counselor behaviors that are influential for chat outcomes. These insights can assist the helpline in their ongoing ambition to improve quality of care.

8.3 Significance to the field

This work underlines the transformative potential of NLP in suicide prevention and other mental health helplines. By automating and enhancing various aspects of helpline operations, from counselor support to administrative insights, NLP technologies pave the way for more resilient and effective suicide prevention services.

The integration of transformer technologies, such as BERT, into suicide prevention helplines has shown great improvements over previous NLP methods, as detailed in Chapters 3, 4, 6 and 7 of this dissertation. These models have been instrumental in processing the complex language patterns and providing contextually embeddings for helpline conversations. This study provides a unique contribution to the field by applying machine learning models to analyze and enhance the efficacy of helpline operations. The findings highlight the potential of NLP systems to not only support counselors in real-time but also offer valuable administrative insights.

Improving the operationality and quality of helplines indirectly also impacts suicide prevention, as helplines are one of the most important interventions to prevent suicide.The higher the quality of care that helplines can provide, the more likely they are to prevent help seekers from becoming victims of suicide.

8.4 Limitations and recommendations for future research

Despite its promising findings, like all research, there are several limitations that should be considered.

In this work, NLP techniques have been applied to online chat conversations, however this is only one part of the helpline. Telephone conversations make up the remainder of the daily interactions with the helpline. With recently improved speech-to-text technologies such as OpenAI Whisper [176], language analysis can also be applied to phone calls. Challenges that must be addressed is mainly what form this analysis will take. Topic modeling and summarisation are approaches that can be easily applied to telephone conversations. However, due to additional overhead from the audio, combined with the quicker pace of a phone call compared to a chat conversation, support tools such as the ones developed in Part I would most likely not be feasible. Labeling of conversations also has to be done in a different manner, as there is no way to easily include an automated questionnaire during a phone call for a person in distress. LLMs might provide potential solutions to these challenges.

Most findings are based on the specific context and demographic of 113, which may limit their generalizability to other settings or populations. To truly understand the impact of NLP methods in helplines, multiple helplines should look to work together in implementing tools for help seekers. While data can not always be shared, techniques, code and the pre-training of models offer multiple avenues for easy collaboration. I highly encourage this collaborative effort in research for suicide prevention through machine learning technologies.

One of the applications of NLP for helplines that have not been explored in this dissertation, but have potential to be very helpful to the helpline are training tools. Simulating conversations can provide a low-stakes environment for counselors to practice, maintain or develop their skills. Furthermore, automated analysis can provide statistics for counselors to more easily review and discuss the conversations they previously had. This is a perfect use-case to make use of the recent developments in LLMs. Open source models can even be used to fine-tune for specific training scenarios through the use of helpline chat data.

Because of these developments, and to make the best use of new AI models, one factor regarding NLP research in mental health that will become increasingly more important will be data and compute infrastructure. Data infrastructure is important because the data is sensitive and at a large scale. This must be well managed to ensure privacy is maintained. Compute infrastructure is important because new NLP machine learning models, especially LLMs, demand a certain amount of GPU memory to opperate, due to increasing model parameters and $O(n^2)$ space requirement of the attention mechanism. Care and consideration should be given to how this will fit into future plans, projects and budgeting. As of the time of writing compute is still a scarce resource and the energy consumption has a negative impact on the environment [177] [175]. Using open source software is also highly encouraged to ensure privacy requirements are met.

However, certain LLM use cases also present significant risks. In Part I, we opted for a retrieval-based system instead of a generative-based one because the retrieved text comes from human-human conversations. A generative approach, on the other hand, requires a degree of trust in its reliability. Studies have been done on using llms for decision support in clinical settings, but found that there are still clear weaknesses [187]. While humans can reason about their decision-making and assume responsibility, a LLM cannot. Before incorporating LLMs into decision-making processes without a thorough understanding of their performance, ethical considerations must be addressed [170]. Moreover, there is no clear consensus and limited evidence-based research on the effects of different helpline intervention approaches. More research is needed to better understand helpline methodologies before introducing

LLMs into this process.

8.5 Concluding remarks

In conclusion, while NLP and machine learning offer substantial advancements for suicide prevention helplines, ongoing research, collaboration, and innovation are essential to fully realize their potential. This dissertation has made strides in demonstrating the impact of these technologies, yet it also highlights the vast opportunities that remain and the potential that is now available, we have only yet scratched the surface. By collaborating and leveraging technologies responsibly and ethically, we can pave the way for more resilient, effective, and compassionate suicide prevention services, turning the power of words into meaningful actions to save lives and offer hope to those in crisis.

Bibliography

1. Turing, A. M. Computer Machinery and Intelligence. *Mind* **LIX,** 433–460 (Oct. 1950).
2. Litman, R. E. Suicide-Prevention Telephone Service. *JAMA: The Journal of the American Medical Association* **192,** 21 (Apr. 1965).
3. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536 (Oct. 1986).
4. Acorn, T. L. & Walden, S. H. *SMART: support management automated reasoning technology for compaq customer service* in *Proceedings of the Fourth Conference on Innovative Applications of Artificial Intelligence* (AAAI Press, San Jose, California, 1992), 3–18.
5. Bishop, C. M. *Neural Networks for Pattern Recognition* (Oxford university press, 1995).
6. Brooke, J. *et al.* SUS-A quick and dirty usability scale. *Usability evaluation in industry* **189,** 4–7 (1996).
7. Hearst, M. A. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics* **23,** 33–64 (1997).
8. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9,** 1735–1780 (Nov. 1997).

9. Goldberg, R. G. & Rosinski, R. R. *Automated natural language understanding customer service system* Accessed: 2020-12-11. Apr. 1999.

10. Isbister, K., Nakanishi, H., Ishida, T. & Nass, C. *Helper agent: designing an assistant for human-human interaction in a virtual meeting space* in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (ACM, Apr. 2000).

11. Loper, E. & Bird, S. *NLTK: the Natural Language Toolkit* in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics* (Association for Computational Linguistics, 2002).

12. Zhang, C. & Chen, T. An active learning framework for content-based information retrieval. *IEEE Transactions on Multimedia* **4,** 260–268 (June 2002).

13. Amrhein, P. C., Miller, W. R., Yahne, C. E., Palmer, M. & Fulcher, L. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of Consulting and Clinical Psychology* **71,** 862–878 (Oct. 2003).

14. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *Journal of Machine learning Research* **3,** 993–1022 (2003).

15. Griffiths, T., Jordan, M., Tenenbaum, J. & Blei, D. *Hierarchical Topic Models and the Nested Chinese Restaurant Process* in *Advances in Neural Information Processing Systems* (eds Thrun, S., Saul, L. & Schölkopf, B.) **16** (MIT Press, 2003).

16. Moyers, T., Martin, T., Catley, D., Harris, K. J. & Ahluwalia, J. S. ASSESSING THE INTEGRITY OF MOTIVATIONAL INTERVIEWING INTERVENTIONS: RELIABILITY OF THE MOTIVATIONAL INTERVIEWING SKILLS CODE. *Behavioural and Cognitive Psychotherapy* **31,** 177–184 (Apr. 2003).

17. Martin, T., Moyers, T., Houck, J. & Christopher, P. Motivational Interviewing Sequential Code for Observing Process Exchanges (MI-SCOPE) coder's manual (2005).

18. Blei, D. M. & Lafferty, J. D. *Dynamic topic models* in *Proceedings of the 23rd international conference on Machine learning - ICML '06* (ACM Press, 2006).

19. Moyers, T. B. & Martin, T. Therapist influence on client language during motivational interviewing sessions. *Journal of Substance Abuse Treatment* **30,** 245–251 (Apr. 2006).

20. Abdi, H. *et al.* Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics* **3,** 2007 (2007).

21. Cialdini, R. B. & Cialdini, R. B. *Influence: The Psychology of Persuasion* (Collins New York, 2007).

22. Gould, M. S., Kalafat, J., HarrisMunfakh, J. L. & Kleinman, M. An Evaluation of Crisis Hotline Outcomes Part 2: Suicidal Callers. *Suicide and Life-Threatening Behavior* **37,** 338–352 (June 2007).

23. Gruber, A., Weiss, Y. & Rosen-Zvi, M. *Hidden Topic Markov Models* in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (eds Meila, M. & Shen, X.) **2** (PMLR, San Juan, Puerto Rico, Mar. 2007), 163–170.

24. Mishara, B. L. *et al.* Which Helper Behaviors and Intervention Styles are Related to Better Short-Term Outcomes in Telephone Crisis Intervention? Results from a Silent Monitoring Study of Calls to the U.S. 1–800-SUICIDE Network. *Suicide and Life-Threatening Behavior* **37,** 308–321 (June 2007).

25. Cristianini, N. & Ricci, E. in *Encyclopedia of Algorithms* 928–932 (Springer US, 2008).

26. Bangor, A., Kortum, P. & Miller, J. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies* **4,** 114–123 (2009).

27. Fukkink, R. & Hermanns, J. Counseling children at a helpline: chatting or calling? *Journal of Community Psychology* **37,** 939–948 (Oct. 2009).

28. Ignacio Madrid, R., Van Oostendorp, H. & Puerta Melguizo, M. C. 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* **25,** 66–75 (Jan. 2009).

29. Rutz, W. & Rihmer, Z. in *Oxford Textbook of Suicidology and Suicide Prevention* 249–256 (Oxford University PressOxford, Mar. 2009).

30. Hong, L. & Davison, B. D. *Empirical study of topic modeling in Twitter* in *Proceedings of the First Workshop on Social Media Analytics* (ACM, July 2010).

31. Weng, J., Lim, E.-P., Jiang, J. & He, Q. *TwitterRank: finding topic-sensitive influential twitterers* in *Proceedings of the third ACM international conference on Web search and data mining* (ACM, Feb. 2010).

32. Du, L., Buntine, W., Jin, H. & Chen, C. Sequential latent Dirichlet allocation. *Knowledge and Information Systems* **31,** 475–503 (June 2011).

33. Gilat, I. & Rosenau, S. Volunteers' perspective of effective interactions with helpline callers: qualitative study. *British Journal of Guidance & Counselling* **39,** 325–337 (Aug. 2011).

34. Sielis, G. A. *et al.* A Context Aware Recommender System for Creativity Support Tools. *Journal of Universal Computer Science* **17,** 1743–1763 (2011).

35. Suler, J. in *Online Counseling* 21–53 (Elsevier, 2011).

36. Uther, W. *et al.* in *Encyclopedia of Machine Learning* 986–987 (Springer US, 2011).

37. Wang, H., Zhang, D. & Zhai, C. *Structural topic model for latent topical structure analysis* in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (2011), 1526–1535.

38. Campello, R. J. G. B., Moulavi, D. & Sander, J. in *Lecture Notes in Computer Science* 160–172 (Springer Berlin Heidelberg, 2013).

39. Dowling, M. & Rickwood, D. Online Counseling and Therapy for Mental Health Problems: A Systematic Review of Individual Synchronous Interventions Using Chat. *Journal of Technology in Human Services* **31,** 1–21 (Jan. 2013).

40. Evans, W. P., Davidson, L. & Sicafuse, L. Someone to Listen: Increasing Youth Help-seeking Behavior Through a Text-Based Crisis Line for Youth. *Journal of Community Psychology* **41,** 471–487 (Mar. 2013).

41. Lundahl, B. *et al.* Motivational interviewing in medical care settings: A systematic review and meta-analysis of randomized controlled trials. *Patient Education and Counseling* **93,** 157–168 (Nov. 2013).

42. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. *Distributed Representations of Words and Phrases and their Compositionality* in *Advances in Neural Information Processing Systems* (eds Burges, C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K.) **26** (Curran Associates, Inc., 2013).

43. Pratt, M. E. *The Future of Volunteers in Crisis Hotline Work* Accessed: 2020-12-11. master's thesis (University of Pittsburgh, Jan. 2013).

44. Yan, X., Guo, J., Lan, Y. & Cheng, X. *A biterm topic model for short texts* in *Proceedings of the 22nd international conference on World Wide Web* (ACM, May 2013).

45. Baroni, M., Dinu, G. & Kruszewski, G. *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)* (Association for Computational Linguistics, 2014).

46. Magill, M. *et al.* The technical hypothesis of motivational interviewing: A meta-analysis of MI's key causal model. *Journal of Consulting and Clinical Psychology* **82,** 973–983 (2014).

47. Schwalbe, C. S., Oh, H. Y. & Zweben, A. Sustaining motivational interviewing: a meta-analysis of training studies. *Addiction* **109,** 1287–1294 (May 2014).

48. Skovholt, T. M. & Trotter-Mathison, M. *The Resilient Practitioner: Burnout Prevention and Self-Care Strategies for Counselors, Therapists, Teachers, and Health Professionals, Second Edition* (Routledge, Apr. 2014).

49. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems* **27** (2014).

50. Yin, J. & Wang, J. *A dirichlet multinomial mixture model-based approach for short text clustering* in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge Discovery and Data Mining* (ACM, Aug. 2014).

51. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67** (2015).

52. Dinakar, K., Chen, J., Lieberman, H., Picard, R. & Filbin, R. *Mixed-Initiative Real-Time Topic Modeling & Visualization for Crisis Counseling* in *Proceedings of the 20th International Conference on Intelligent User Interfaces* (ACM, Mar. 2015).

53. Gould, M. S. *et al.* Helping Callers to the National Suicide Prevention Lifeline Who Are at Imminent Risk of Suicide: Evaluation of Caller Risk Profiles and Interventions Implemented. *Suicide and Life-Threatening Behavior* **46,** 172–190 (Aug. 2015).

54. Gunaratne, J. & Nov, O. in *Lecture Notes in Computer Science* 205–216 (Springer International Publishing, 2015).

55. Kumar, M., Dredze, M., Coppersmith, G. & De Choudhury, M. *Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides* in *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15* (ACM Press, 2015).

56. Li, J., Luong, M.-T. & Jurafsky, D. A Hierarchical Neural Autoencoder for Paragraphs and Documents. arXiv: 1506.01057 [cs.CL] (2015).

57. Lindson-Hawley, N., Thompson, T. P. & Begh, R. Motivational interviewing for smoking cessation. *Cochrane Database of Systematic Reviews* (Mar. 2015).

58. Pennebaker, J. W., Boyd, R. L., Jordan, K. & Blackburn, K. The Development and Psychometric Properties of LIWC2015 (2015).

59. Röder, M., Both, A. & Hinneburg, A. *Exploring the Space of Topic Coherence Measures* in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (ACM, Feb. 2015).

60. Wang, C., Blei, D. & Heckerman, D. Continuous Time Dynamic Topic Models. arXiv: 1206.3298 [cs.IR] (2015).

61. Achakulvisut, T., Acuna, D. E., Ruangrong, T. & Kording, K. Science Concierge: A Fast Content-Based Recommendation System for Scientific Publications. *PLOS ONE* **11** (ed van den Besselaar, P.) e0158423 (July 2016).

62. Althoff, T., Clark, K. & Leskovec, J. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics* **4,** 463–476 (Dec. 2016).

63. Mokkenstorm, J. K. *et al.* Evaluation of the 113Online Suicide Prevention Crisis Chat Service: Outcomes, Helper Behaviors and Comparison to Telephone Hotlines. *Suicide and Life-Threatening Behavior* **47,** 282–296 (Aug. 2016).

64. Palacio, A. *et al.* Motivational Interviewing Improves Medication Adherence: a Systematic Review and Meta-analysis. *Journal of General Internal Medicine* **31,** 929–940 (May 2016).

65. Palangi, H. *et al.* Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **24,** 694–707 (Apr. 2016).

66. Ribeiro, M. T., Singh, S. & Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv: 1602.04938 [cs.LG] (2016).

67. Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C. & Srikumar, V. A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of Substance Abuse Treatment* **65,** 43–50 (June 2016).

68. Tian, F., Gao, B., He, D. & Liu, T.-Y. Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves. arXiv: 1604.02038 [cs.LG] (2016).

69. Turecki, G. & Brent, D. A. Suicide and suicidal behaviour. *The Lancet* **387,** 1227–1239 (Mar. 2016).

70. Wang, T., Huang, Z. & Gan, C. On mining latent topics from healthcare chat logs. *Journal of Biomedical Informatics* **61,** 247–259 (June 2016).

71. Yang, Z. *et al. Hierarchical Attention Networks for Document Classification* in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, 2016).

72. Arora, S., Liang, Y. & Ma, T. *A Simple but Tough-to-Beat Baseline for Sentence Embeddings* in *International Conference on Learning Representations* (2017).

73. Cameron, G. *et al. Towards a chatbot for digital counselling* in *Electronic Workshops in Computing* (BCS Learning & Development, 2017).

74. DiClemente, C. C., Corno, C. M., Graydon, M. M., Wiprovnick, A. E. & Knoblach, D. J. Motivational interviewing, enhancement, and brief interventions over the last decade: A review of reviews of efficacy and effectiveness. *Psychology of Addictive Behaviors* **31,** 862–887 (Dec. 2017).

75. Franklin, J. C. *et al.* Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin* **143,** 187–232 (2017).

76. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. arXiv: 1705.07874 [cs.AI] (2017).

77. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* **2,** 205 (Mar. 2017).

78. Pérez-Rosas, V. *et al. Predicting Counselor Behaviors in Motivational Interviewing Encounters* in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (eds Lapata, M., Blunsom, P. & Koller, A.) (Association for Computational Linguistics, Valencia, Spain, Apr. 2017), 1128–1137.

79. Phillips, R., Hogden, A. & Greenfield, D. in *Promoting Self-Management of Chronic Health Conditions* (ed Martz, E.) 126–144 (Oxford University Press, July 2017).

80. Qiu, M. *et al. AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine* in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (Association for Computational Linguistics, 2017).

81. Srivastava, A. & Sutton, C. Autoencoding Variational Inference For Topic Models. arXiv: 1703.01488 [stat.ML] (2017).

82. Walsh, C. G., Ribeiro, J. D. & Franklin, J. C. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science* **5,** 457–469 (Apr. 2017).

83. Coppersmith, G., Leary, R., Crutchley, P. & Fine, A. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights* **10,** 117822261879286 (Jan. 2018).

84. Frost, H. *et al.* Effectiveness of Motivational Interviewing on adult behaviour change in health and social care settings: A systematic review of reviews. *PLOS ONE* **13** (ed Moitra, E.) e0204890 (Oct. 2018).

85. Hallgren, K. A. *et al.* Variability in motivational interviewing adherence across sessions, providers, sites, and research contexts. *Journal of Substance Abuse Treatment* **84,** 30–41 (Jan. 2018).

86. Hasan, M. *et al.* Identifying Effective Motivational Interviewing Communication Sequences Using Automated Pattern Analysis. *Journal of Healthcare Informatics Research* **3,** 86–106 (Oct. 2018).

87. Li, X. *et al.* Filtering out the noise in short text topic modeling. *Information Sciences* **456,** 83–96 (Aug. 2018).

88. Li, X. *et al.* Relational Biterm Topic Model: Short-Text Topic Modeling using Word Embeddings. *The Computer Journal* **62,** 359–372 (May 2018).

89. Magill, M. *et al.* A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of Consulting and Clinical Psychology* **86,** 140–157 (Feb. 2018).

90. Nieuwenhuijse, A. *COOSTO Dutch Word2Vec Model* Accessed: 2019-03-20. July 2018.

91. O'Connor, R. C. & Kirtley, O. J. The integrated motivational–volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373,** 20170268 (July 2018).

92. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., *et al. Improving language understanding by generative pre-training* Accessed: 2024-07-11. June 2018.

93. Yang, Y. *et al. Learning Semantic Textual Similarity from Conversations* in *Proceedings of The Third Workshop on Representation Learning for NLP* (Association for Computational Linguistics, 2018).

94. Ailem, M., Zhang, B. & Sha, F. Topic Augmented Generator for Abstractive Summarization. arXiv: 1908.07026 [cs.LG] (2019).

95. De Vries, W. *et al.* BERTje: A Dutch BERT Model. arXiv: 1912.09582 [cs.CL] (2019).

96. Demasi, O., Hearst, M. A. & Recht, B. *Towards augmenting crisis counselor training by improving message retrieval* in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (Association for Computational Linguistics, 2019).

97. Derrar, H. M. *Clustering for the automatic annotation of customer service chat messages* Accessed: 2020-12-11. master's thesis (Tampere University, Jan. 2019).

98. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv: 1810.04805 [cs.CL] (2019).

99. Fiske, A., Henningsen, P. & Buyx, A. Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy. *Journal of Medical Internet Research* **21,** e13216 (May 2019).

100. Gao, S. *et al.* Classifying cancer pathology reports with hierarchical self-attention networks. *Artificial Intelligence in Medicine* **101,** 101726 (Nov. 2019).

101. Idalski Carcone, A. *et al.* Developing Machine Learning Models for Behavioral Coding. *Journal of Pediatric Psychology* **44,** 289–299 (Jan. 2019).

102. Imel, Z. E. *et al.* Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy* **56,** 318–328 (June 2019).

103. Jain, S. & Wallace, B. C. *Attention is not Explanation* in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J., Doran, C. & Solorio, T.) (Association for Computational Linguistics, Minneapolis, Minnesota, June 2019), 3543–3556.

104. Kulasinghe, S. *et al. AI Based Depression and Suicide Prevention System* in *2019 International Conference on Advancements in Computing (ICAC)* (IEEE, Dec. 2019).

105. Liu, Y. *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv: 1907.11692 [cs.CL] (2019).

106. Lundahl, B. *et al.* Motivational interviewing adherence tools: A scoping review investigating content validity. *Patient Education and Counseling* **102,** 2145–2155 (Dec. 2019).

107. Rashid, J., Shah, S. M. A. & Irtaza, A. Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management* **56,** 102060 (Nov. 2019).

108. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv: 1908.10084 [cs.CL] (2019).

109. Salmi, S. *Context-based recommender system to provide cognitive support to online chat counsellors in the Helpline of 113 Suicide Prevention.* Accessed: 2020-12-11. master's thesis (Delft University of Technology, Nov. 2019).

110. Serrano, S. & Smith, N. A. Is Attention Interpretable? arXiv: 1906.03731 [cs.CL] (2019).

111. Wallace, E. *et al.* AllenNLP Interpret: A Framework for Explaining Predictions of NLP Models. arXiv: 1909.09251 [cs.CL] (2019).

112. Wiegreffe, S. & Pinter, Y. Attention is not not Explanation. arXiv: 1908.04626 [cs.CL] (2019).

113. Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M. & Househ, M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research* **22,** e16021 (July 2020).

114. Angelov, D. Top2Vec: Distributed Representations of Topics. arXiv: 2008.09470 [cs.CL] (2020).

115. Bastings, J. & Filippova, K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? arXiv: 2010.05607 [cs.CL] (2020).

116. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The Long-Document Transformer. arXiv: 2004.05150 [cs.CL] (2020).

117. Crone, E. A. & Fuligni, A. J. Self and Others in Adolescence. *Annual Review of Psychology* **71,** 447–469 (Jan. 2020).

118. Delobelle, P., Winters, T. & Berendt, B. RobBERT: a Dutch RoBERTa-based Language Model. arXiv: 2001.06286 [cs.CL] (2020).

119. Dieng, A. B., Ruiz, F. J. R. & Blei, D. M. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics* **8,** 439–453 (Dec. 2020).

120. Gunnell, D. *et al.* Suicide risk and prevention during the COVID-19 pandemic. *The Lancet Psychiatry* **7,** 468–471 (June 2020).

121. Hurlocker, M. C., Madson, M. B. & Schumacher, J. A. Motivational interviewing quality assurance: A systematic review of assessment tools across research contexts. *Clinical Psychology Review* **82,** 101909 (Dec. 2020).

122. Madeira, T. *et al. A Framework to Assist Chat Operators of Mental Healthcare Services* in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)* (Association for Computational Linguistics, 2020).

123. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: `1802.03426 [stat.ML]` (2020).

124. Meng, Y. *et al. Discriminative Topic Mining via Category-Name Guided Text Embedding* in *Proceedings of The Web Conference 2020* (ACM, Apr. 2020).

125. O'Connor, R. C. *et al.* Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 Mental Health & Wellbeing study. *The British Journal of Psychiatry* **218,** 326–333 (Oct. 2020).

126. Padmanathan, P., Bould, H., Winstone, L., Moran, P. & Gunnell, D. Social media use, economic recession and income inequality in relation to trends in youth suicide in high-income countries: a time trends analysis. *Journal of Affective Disorders* **275,** 58–65 (Oct. 2020).

127. Pierce, M. *et al.* Mental health before and during the COVID-19 pandemic: a longitudinal probability sample survey of the UK population. *The Lancet Psychiatry* **7,** 883–892 (Oct. 2020).

128. Reger, M. A., Stanley, I. H. & Joiner, T. E. Suicide Mortality and Coronavirus Disease 2019—A Perfect Storm? *JAMA Psychiatry* **77,** 1093 (Nov. 2020).

129. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv: 1910.01108 [cs.CL] (2020).

130. Sharma, A., Miner, A. S., Atkins, D. C. & Althoff, T. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. arXiv: 2009.08441 [cs.CL] (2020).

131. Sia, S., Dalmia, A. & Mielke, S. J. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! arXiv: 2004.14914 [cs.CL] (2020).

132. Wang, C. *et al.* Immediate Psychological Responses and Associated Factors during the Initial Stage of the 2019 Coronavirus Disease (COVID-19) Epidemic among the General Population in China. *International Journal of Environmental Research and Public Health* **17,** 1729 (Mar. 2020).

133. Wang, Z. *et al. Friendly Topic Assistant for Transformer Based Abstractive Summarization* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, 2020).

134. World Health Organisation. *WHO Director-General's opening remarks at the media briefing on COVID-19* Accessed: 2020-09-30. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19%2D%2D-11-march-2020.

135. Yang, Z. *et al.* XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv: 1906.08237 [cs.CL] (2020).

136. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (Feb. 2020).

137. Ahmadi, A. *et al.* A Systematic Review of Machine Learning for Assessment and Feedback of Treatment Fidelity. *Psychosocial Intervention* **30,** 139–153 (July 2021).

138. Bianchi, F., Terragni, S. & Hovy, D. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. arXiv: `2004.03974 [cs.CL]` (2021).

139. CANS. *CANS Duiding februari: Geen toename in suïcides* Accessed: 2021-02-22. `https://www.113.nl/actueel/cans-duiding-februari-geen-toename-suicides`.

140. Flemotomos, N. *et al.* Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods* **54,** 690–711 (Aug. 2021).

141. Gould, M. S. *et al.* National Suicide Prevention Lifeline crisis chat interventions: Evaluation of chatters' perceptions of effectiveness. *Suicide and Life-Threatening Behavior* **51,** 1126–1137 (July 2021).

142. Otter, D. W., Medina, J. R. & Kalita, J. K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems* **32,** 604–624 (Feb. 2021).

143. Pirkis, J. *et al.* Suicide trends in the early months of the COVID-19 pandemic: an interrupted time-series analysis of preliminary data from 21 countries. *The Lancet Psychiatry* **8,** 579–588 (July 2021).

144. Salmi, S., Mérelle, S. Y. M., Gilissen, R. & Brinkman, W. P. Content-Based Recommender Support System for Counselors in a Suicide Prevention Chat Helpline: Design and Evaluation Study. *Journal of Medical Internet Research* **23,** e21690 (Jan. 2021).

145. Salmi, S., Mérelle, S. Y. M., Gilissen, R. & Brinkman, W. P. *Data and analysis for the publication: Content-Based Recommender Support System for Counselors in a Suicide Prevention Chat*

Helpline: Design and Evaluation Study https://data.4tu.nl/articles/_/13366739.

146. Schreuders, E., Braams, B. R., Crone, E. A. & Güroğlu, B. Friendship stability in adolescence is associated with ventral striatum responses to vicarious rewards. *Nature Communications* **12** (Jan. 2021).

147. Solbiati, A. *et al.* Unsupervised Topic Segmentation of Meetings with BERT Embeddings. arXiv: 2106.12978 [cs.LG] (2021).

148. Tavabi, L. *et al. Analysis of Behavior Classification in Motivational Interviewing* in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access* (Association for Computational Linguistics, 2021).

149. Vahdat, A. & Kautz, J. NVAE: A Deep Hierarchical Variational Autoencoder. arXiv: 2007.03898 [stat.ML] (2021).

150. Wang, K., Reimers, N. & Gurevych, I. *TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning* in *Findings of the Association for Computational Linguistics: EMNLP 2021* (Association for Computational Linguistics, 2021).

151. Xing, L. & Carenini, G. Improving Unsupervised Dialogue Topic Segmentation with Utterance-Pair Coherence Scoring. arXiv: 2106.06719 [cs.CL] (2021).

152. Côté, L.-P. & Mishara, B. L. Effect of helping suicidal people using text messaging: An evaluation of effects and best practices of the Canadian suicide prevention Service's text helpline. *Suicide and Life-Threatening Behavior* **52,** 1140–1148 (Aug. 2022).

153. Fischer, K. *et al.* Internalizing problems before and during the COVID-19 pandemic in independent samples of Dutch children and adolescents with and without pre-existing mental health problems. *European Child &amp; Adolescent Psychiatry* **32,** 1873–1883 (May 2022).

154. Geigle, G., Pfeiffer, J., Reimers, N., Vulić, I. & Gurevych, I. Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. *Transactions of the Association for Computational Linguistics* **10** (eds Roark, B. & Nenkova, A.) 503–521 (2022).

155. Gould, M. S. *et al.* Crisis text-line interventions: Evaluation of texters' perceptions of effectiveness. *Suicide and Life-Threatening Behavior* **52,** 583–595 (May 2022).

156. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv: `2203.05794 [cs.CL]` (2022).

157. Ilias, L. & Askounis, D. Explainable Identification of Dementia From Transcripts Using Transformer Networks. *IEEE Journal of Biomedical and Health Informatics* **26,** 4153–4164 (Aug. 2022).

158. Janssen, W., Raak, J. v., Lucht, Y. v. d., Ballegooijen, W. v. & Mérelle, S. Can Outcomes of a Chat-Based Suicide Prevention Helpline Be Improved by Training Counselors in Motivational Interviewing? A Non-randomized Controlled Trial. *Frontiers in Digital Health* **4** (June 2022).

159. Kirtley, O. J., van Mens, K., Hoogendoorn, M., Kapur, N. & de Beurs, D. Translating promise into practice: a review of machine learning in suicide research and prevention. *The Lancet Psychiatry* **9,** 243–252 (Mar. 2022).

160. Meng, Y., Zhang, Y., Huang, J., Zhang, Y. & Han, J. *Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations* in *Proceedings of the ACM Web Conference 2022* (ACM, Apr. 2022).

161. Michalopoulou, M. *et al.* Effectiveness of Motivational Interviewing in Managing Overweight and Obesity: A Systematic Review and Meta-analysis. *Annals of Internal Medicine* **175,** 838–850 (June 2022).

162. Pas, E. T., Borden, L., Debnam, K. J., De Lucia, D. & Bradshaw, C. P. Exploring profiles of coaches' fidelity to Double Check's Motivational Interviewing-embedded coaching: Outcomes associated with fidelity. *Journal of School Psychology* **92,** 285–298 (June 2022).

163. Pisani, A. R. *et al.* Individuals who text crisis text line: Key characteristics and opportunities for suicide prevention. *Suicide and Life-Threatening Behavior* **52,** 567–582 (May 2022).

164. Saiyed, A. *et al.* Technology-Assisted Motivational Interviewing: Developing a Scalable Framework for Promoting Engagement with Tobacco Cessation Using NLP and Machine Learning. *Procedia Computer Science* **206,** 121–131 (2022).

165. Salmi, S. & Brinkman, W. P. *Evaluation inspiration support system for suicide crisis counseling* `https://osf.io/9gu2y/`.

166. Salmi, S., Mérelle, S. Y. M., Gilissen, R., van der Mei, R. D. & Bhulai, S. Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID-19 pandemic: in-depth analysis using encoder representations from transformers. *BMC Public Health* **22** (Mar. 2022).

167. Sedlakova, J. & Trachsel, M. Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent? *The American Journal of Bioethics* **23,** 4–13 (Apr. 2022).

168. Wu, Z. *et al. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues* in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, May 2022).

169. Zhao, W. X., Liu, J., Ren, R. & Wen, J.-R. Dense Text Retrieval based on Pretrained Language Models: A Survey. arXiv: `2211.14876 [cs.IR]` (2022).

170. Cabrera, J., Loyola, M. S., Magaña, I. & Rojas, R. in *Lecture Notes in Computer Science* 313–326 (Springer Nature Switzerland, 2023).

171. De Freitas, J., Agarwal, S., Schmitt, B. & Haslam, N. Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour* **7,** 1845–1854 (Nov. 2023).

172. Farhat, F. ChatGPT as a Complementary Mental Health Resource: A Boon or a Bane. *Annals of Biomedical Engineering* (July 2023).

173. Gao, H. *et al.* Unsupervised Dialogue Topic Segmentation with Topic-aware Utterance Representation. arXiv: `2305.02747 [cs.CL]` (2023).

174. Hsu, S.-L. *et al.* Helping the Helper: Supporting Peer Counselors via AI-Empowered Practice and Feedback. arXiv: `2305.08982 [cs.HC]` (2023).

175. Luccioni, S. *The Mounting Human and Environmental Costs of Generative AI* Accessed: 2024-07-01. `https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/` (2023).

176. Radford, A. *et al. Robust Speech Recognition via Large-Scale Weak Supervision* in *Proceedings of the 40th International Conference on Machine Learning* (eds Krause, A. *et al.*) **202** (PMLR, July 2023), 28492–28518.

177. Saul, J. & Bass, D. *Artificial Intelligence Is Booming—So Is Its Carbon Footprint* Accessed: 2024-07-01. `https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure#xj4y7vzkg` (2023).

178. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* **5,** 46–57 (Jan. 2023).

179. Van der Burgt, M. C. A., Mérelle, S. Y. M., Beekman, A. T. F. & Gilissen, R. The Impact of COVID-19 on the Suicide Prevention Helpline in The Netherlands. *Crisis* **44,** 285–291 (July 2023).

180. Vaswani, A. *et al.* Attention Is All You Need. arXiv: 1706.03762 [cs.CL] (2023).

181. World Health Organisation. *Suicide* Accessed: 2024-04-17. https://www.who.int/news-room/fact-sheets/detail/suicide.

182. Wutz, M., Hermes, M., Winter, V. & Köberlein-Neu, J. Factors Influencing the Acceptability, Acceptance, and Adoption of Conversational Agents in Health Care: Integrative Review. *Journal of Medical Internet Research* **25,** e46548 (Sept. 2023).

183. Zech, J. M. *et al.* An Integrative Engagement Model of Digital Psychotherapy: Exploratory Focus Group Findings. *JMIR Formative Research* **7,** e41428 (Apr. 2023).

184. 113 Zelfmoordpreventie. *Homepage of 113 Zelfmoordpreventie* https://113.nl (2024).

185. 113 Zelfmoordpreventie. *Sterk groeiend aantal hulpgesprekken met 113 Zelfmoordpreventie* Accessed: 2024-07-11. https://www.113.nl/actueel/sterk-groeiend-aantal-hulpgesprekken-met-113-zelfmoordpreventie (2024).

186. Salmi, S., van der Mei, R. D., Mérelle, S. Y. M. & Bhulai, S. Topic modeling for conversations for mental health helplines with utterance embedding. *Telematics and Informatics Reports* **13,** 100126 (Mar. 2024).

187. Sandmann, S., Riepenhausen, S., Plagwitz, L. & Varghese, J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nature Communications* **15** (Mar. 2024).

188. Pellemans, M., Salmi, S., Mérelle, S. Y. M., Janssen, W. C. & van der Mei, R. D. Automated Behavioral Coding to Enhance the Effectiveness of Motivational Interviewing in a Chat-Based Suicide Prevention Helpline. *Journal of Medical Internet Research.* Accepted for publication.

189. Salmi, S., Mérelle, S., Gilissen, R., van der Mei, R. D. & Bhulai, S. The most effective interventions during online suicide prevention chats: Machine Learning Study. Under review.

190. Salmi, S. *et al.* Real-time assistance in suicide prevention helplines using a deep learning-based recommender system: a randomized controlled trial. Submitted for publication.

Summary

This dissertation explores the integration of Natural Language Processing and machine learning techniques to enhance the effectiveness and efficiency of suicide prevention helplines. By developing and evaluating AI-driven support tools, this research aims to provide real-time assistance to counselors, improve the quality of helpline services, and gain deeper insights into the challenges faced by individuals in crisis.

Chapter 2 outlines the design and development of a content-based recommender support system aimed at assisting counselors in a suicide prevention helpline. The primary motivation was to address the cognitive and emotional challenges faced by counselors during intense chat conversations. By implementing a system that provides real-time suggestions based on previous successful interactions, the goal was to reduce response time and alleviate mental fatigue. The evaluation of this tool in a simulated environment showed.

Building on the support tool developed in Chapter 2, Chapter 3 presents an evaluation of the effectiveness through a randomized controlled trial. The recommender system was improved using a deep-learning method and tested in real-time counseling sessions to assess its impact on counselors' self-efficacy and the quality of their responses. While the tool did not significantly improve self-efficacy scores, it was frequently used in longer, more complex conversations, suggesting its utility in challenging situations. The study highlighted the feasibility of integrating AI-assisted tools in helpline services, paving the way for future enhancements.

Chapter 4 focuses on evaluating various topic modeling methods to analyze conversation data from mental health helplines. Traditional methods like Latent Dirichlet Allocation were compared with newer techniques such as BERTopic, which leverages sentence embeddings for better context capture. The study found that BERTopic outperformed other methods in terms of topic coherence and interpretability, especially when applied to short, context-dependent texts typical of helpline conversations. The insights gained from this analysis can help helplines better understand the issues faced by their clients and improve their services.

The aim of Chapter 5 was to identify changes in conversation topics on a suicide prevention helpline during the COVID-19 pandemic. Using BERTopic, the study analyzed chat data before and after the lockdown measures were implemented. The results indicated significant shifts in conversation topics, with an increased mention expressions of gratitude towards counselors. There were decreases in mentions of specific suicide plans, however, specifically helpseekers who lived alone showed a worrying increase in plans for suicide. These findings underscore the impact of the pandemic on mental health and the potential for monitoring of helpline conversations.

Chapter 6 explores the use of AI models to classify counselor and client behaviors in the context of Motivational Interviewing (MI) during helpline chats. By training models on a coded dataset of MI sessions, the study aimed to automate the identification of effective counseling techniques. The deep learning model BERTje showed high accuracy in classifying MI behaviors, indicating its potential as a tool for providing real-time feedback to counselors. This automation could enhance the quality of MI in online helplines, ensuring that counselors adhere to best practices and improve client outcomes .

Finally, Chapter 7 investigates which counselor utterances contribute to positive outcomes for help seekers using deep learning models. By analyzing chat logs and help seeker self-assessments,

the study identified key behaviors of counselors that positively or negatively affected client wellbeing. Positive affirmations and expressing involvement were linked to improved scores, while the use of macros and premature conversation endings had negative effects. The study highlighted the potential of machine learning to provide actionable insights for training counselors and enhancing the effectiveness of helpline conversations.