The most effective interventions during online suicide prevention chats: Machine Learning Study

Salim Salmi, Saskia Mérelle, Renske Gilissen, Rob van der Mei, Sandjai Bhulai

Submitted to: JMIR Mental Health
on: February 14, 2024

Table of Contents

The most effective interventions during online suicide prevention chats: Machine Learning Study

Salim Salmi[1] MSc; Saskia Mérelle[1] PhD; Renske Gilissen[1] PhD; Rob van der Mei[2] PhD; Sandjai Bhulai[3] PhD

[1]113 Suicide Prevention Amsterdam NL
[2]Stochastics Centrum Wiskunde & Informatica Amsterdam NL
[3]Mathematics Vrije Universiteit Amsterdam Amsterdam NL

Corresponding Author:
Salim Salmi MSc
113 Suicide Prevention
Paasheuvelweg 25
Amsterdam
NL

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.

Objective: We trained a machine learning classification model and identify which counsellor utterances have the most impact on its outputs.

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 the suicide prevention helpline in the Netherlands (113 Suicide Prevention). Machine learning 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 machine learning 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.

Conclusions: 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.

(JMIR Preprints 14/02/2024:57362)
DOI: https://doi.org/10.2196/preprints.57362

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

The most effective interventions during online suicide prevention chats: Machine Learning Study

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.
Objective: We trained a machine learning classification model and identify which counsellor utterances have the most impact on its outputs.
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 the suicide prevention helpline in the Netherlands (113 Suicide Prevention). Machine learning 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 machine learning 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.
Conclusions: 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.

Keywords: Classification; Interpretable AI; Conversations; Suicide prevention; BERT

Introduction

Worldwide, helplines have been set up to answer thousands of people with suicidal thoughts every day. With technology advancing and the internet having become a big presence in daily life, helplines can now often also be contacted online through chat services.
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 [1, 2]. Furthermore, Gould et al.[3] examined call reports of help seekers calling helplines in the National Suicide Prevention Lifeline network. In studies by Mokkenstorm et al.[4] and Mishara et al. [5] helpline chat logs were annotated and analyzed 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.
In recent years, significant advancements have been made in the field of natural-language analysis. Deep learning models such as transformers enabled more effective use of big data [6]. Furthermore, BERT models [7] made transfer learning a more viable practice. This provided an opportunity to use

these methods to do a large-scale analysis of helpline chat data.

113 Suicide Prevention, the national suicide prevention helpline in the Netherlands, uses pre- and post-conversation questionnaires to assess the help seeker's mental wellbeing [8], 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.

Related Work

Many state-of-the-art language models that have been developed in recent years rely on transformers. First introduced by Vaswani et al. [6], transformers leverage the self-attention mechanism to create long-range connections in a sequential input. This mechanism uses scaled dot-product attention (1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right)V \qquad (1)$$

The projections of the input sequence Q and K are used to compute a weighted average of a final projection V. To prevent the weights from getting too large, in turn causing the gradients to become too small, they are scaled by the dimension d_k of the input sequence. In the original paper, the transformer was developed as an encoder-decoder network for the task of language translation. Devlin et al. [7] 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 [9]. 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 [10] found that attention weights only noisily predict importance.
Jain and Wallace [11] argue attention weights do not provide explanations, while Wiegreffe and Pinter [12] 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 [13] [14]. However, even saliency maps have limitations [14]. Local Interpretable Model-Agnostic Explanations (LIME) [15] 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 [16], 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 [17]. 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, 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. [18] 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 [19] used LIME to find influential words of BERT classifications of Dementia transcripts.

Methods

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.

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 counsellor. 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.

Chat message embedding

The individual messages were embedded using a pre-trained RoBERTa network called RobBERT[20]. 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 $C \in R^{b \times l \times d}$, where l is the length of the sequence, d the embedding size of the pre-trained network, and b the batch size.

Weighted average

To improve explainability, we used a simpler adaptation of the attention mechanism. The weighted average is described in (2).

$$\text{Weighted Average}(C) = \text{softmax}\left(\left(C W_k^T + b_k\right)^T\right)\left(C W_v^T + b_v\right) \quad (2)$$

Here $C \in R^{n \times d}$ is the matrix of input embeddings with embedding dimension d, from messages belonging to a conversation of length n. $W_k \in R^{1 \times d}$ and $W_v \in R^{d \times d}$ 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 counsellor 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 counsellor message and one for only the help seeker messages. This resulted in the weighted average only being an average of either the counsellor or help seeker. The counsellor and help seeker each had the same number of heads.

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 long-short term memory (LSTM) [21] 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.

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 [22]. Finally, each embedded conversation was trained on a support vector machine [23].
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.

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 important to the result of the conversation. As a second technique, we applied LIME [15] to the models. For this approach, we left out counsellor messages one at a time to compute the difference in loss. A larger difference indicates more importance to the classification.

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).

Results

Model performance

Table 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

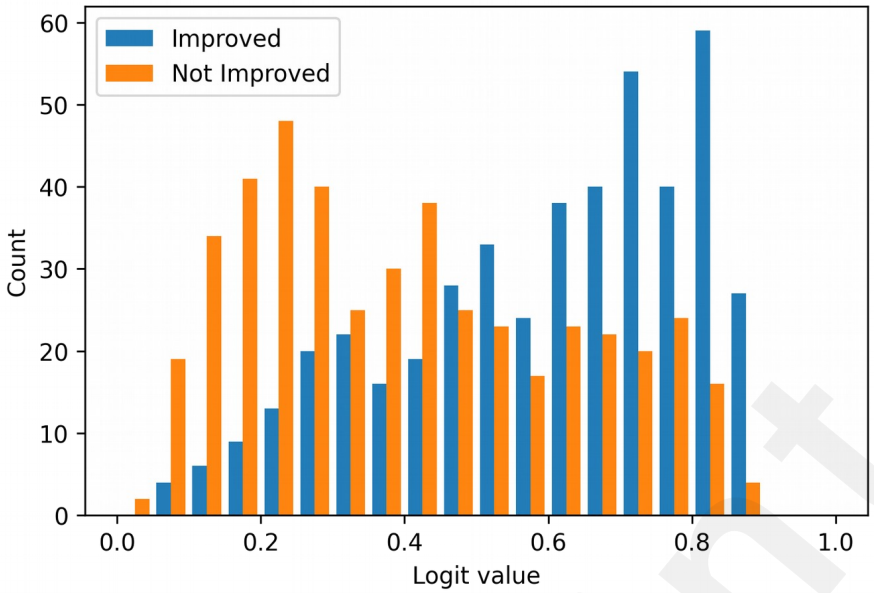
Table 1 shows the performance scores on a held-out test set. The Hierarchical Weighted Average model was the best performing model with an accuracy of 0.683 and the highest F1 Score of 0.688.

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.

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

Figure 1: Histogram of logits for the Hierarchical Weighted model



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 counsellor 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 counsellor 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 counsellor would ask open ended questions, as well as show respect for the autonomy of the help seeker by asking what they wanted to do.

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.

The research by Mishara et al. [5] 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, Cˆote and Mishara [24] 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 [25] 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.

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 counsellor 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.

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.

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 counsellors can employ. Including positive reinforcement more regularly in their responses might be beneficial to helpline conversations.

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.

Acknowledgements

We would like to thank the senior psychologists Mirjam van Driel and Maryke Geerdink for their interpretation of the impactful messages.

SS had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Study concept and design: All authors. Acquisition, analysis, or interpretation of the data: SS. Drafting of the manuscript: SS. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: SS. Administrative, technical or material support: RG, RM. Study supervision: SB, SM. The authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
LIME: Local Interpretable Model-Agnostic Explanations
LSTM: long-short term memory
NLP: Natural Language Processing
RoBERTa: Robustly optimized BERT pretraining approach

References

1. Gould MS, Pisani A, Gallo C, Ertefaie A, Harrington D, Kelberman C, Green S. Crisis text-line interventions: Evaluation of texters perceptions of effectiveness. Suicide and Life-Threatening Behavior Wiley; 2022 May;52(3):583–595. doi: 10.1111/sltb.12873

2. Pisani AR, Gould MS, Gallo C, Ertefaie A, Kelberman C, Harrington D, Weller D, Green S. Individuals who text crisis text line: Key characteristics and opportunities for suicide prevention. Suicide and Life-Threatening Behavior Wiley; 2022 May;52(3):567–582. doi: 10.1111/sltb.12872

3. Gould MS, Lake AM, Munfakh JL, Galfalvy H, Kleinman M, Williams C, Glass A, McKeon R. 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 Wiley; 2015 Aug;46(2):172–190. doi: 10.1111/sltb.12182

4. Mokkenstorm JK, Eikelenboom M, Huisman A, Wiebenga J, Gilissen R, Kerkhof AJFM, Smit JH. Evaluation of the 113Online suicide prevention crisis chat service: Outcomes, helper behaviors and comparison to telephone hotlines. Suicide and Life-Threatening Behavior Wiley; 2016 Aug;47(3):282–296. doi: 10.1111/sltb.12286

5. Mishara BL, Chagnon F, Daigle M, Balan B, Raymond S, Marcoux I, Bardon C, Campbell JK, Berman A. 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. 1800-SUICIDE network. Suicide and Life-Threatening Behavior Wiley; 2007 Jun;37(3):308–321. doi: 10.1521/suli.2007.37.3.308

6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. arXiv; 2017. doi: 10.48550/ARXIV.1706.03762

7. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv; 2018. doi: 10.48550/ARXIV.1810.04805

8. Janssen W, Raak J van, Lucht Y van der, Ballegooijen W van, 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 Frontiers Media SA; 2022 Jun;4. doi: 10.3389/fdgth.2022.871841

9. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv; 2019. doi: 10.48550/ARXIV.1907.11692

10. Serrano S, Smith NA. Is attention interpretable? arXiv; 2019. doi: 10.48550/ARXIV.1906.03731

11. Jain S, Wallace BC. Proceedings of the 2019 conference of the north Association for Computational Linguistics; 2019. doi: 10.18653/v1/n19-1357

12. Wiegreffe S, Pinter Y. Attention is not not explanation. CoRR 2019;abs/1908.04626. Available from: http://arxiv.org/abs/1908.04626

13. Wallace E, Tuyls J, Wang J, Subramanian S, Gardner M, Singh S. AllenNLP interpret: A framework for explaining predictions of NLP models. arXiv; 2019. doi: 10.48550/ARXIV.1909.09251

14. Bastings J, Filippova K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? arXiv; 2020; doi: 10.48550/ARXIV.2010.05607

15. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining New York, NY, USA: Association for Computing Machinery; 2016. p. 1135–1144. doi: 10.1145/2939672.2939778

16. Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv; 2020. doi: 10.48550/ARXIV.2004.05150

17. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies Association for Computational Linguistics; 2016. doi: 10.18653/v1/n16-1174

18. Gao S, Qiu JX, Alawad M, Hinkle JD, Schaefferkoetter N, Yoon H-J, Christian B, Fearn PA, Penberthy L, Wu X-C, Coyle L, Tourassi G, Ramanathan A. Classifying cancer pathology reports with hierarchical self-attention networks. Artificial Intelligence in Medicine Elsevier BV; 2019 Nov;101:101726. doi: 10.1016/j.artmed.2019.101726

19. Ilias L, Askounis D. Explainable identification of dementia from transcripts using transformer networks. IEEE Journal of Biomedical and Health Informatics Institute of Electrical; Electronics Engineers (IEEE); 2022 Aug;26(8):4153–4164. doi: 10.1109/jbhi.2022.3172479

20. Delobelle P, Winters T, Berendt B. RobBERT: A dutch RoBERTa-based language model. arXiv; 2020. doi: 10.48550/ARXIV.2001.06286

21. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation MIT Press - Journals; 1997 Nov;9(8):1735–1780. doi: 10.1162/neco.1997.9.8.1735

22. Uther W, Mladenić D, Ciaramita M, Berendt B, Kołcz A, Grobelnik M, Mladenić D, Witbrock M, Risch J, Bohn S, Poteet S, Kao A, Quach L, Wu J, Keogh E, Miikkulainen R, Flener P, Schmid U, Zheng F, Webb GI, Nijssen S. TFIDF. Encyclopedia of machine learning Springer US; 2011. p. 986–987. doi: 10.1007/978-0-387-30164-8_832

23. Cristianini N, Ricci E. Support vector machines. Encyclopedia of algorithms Springer US; 2008. p. 928–932. doi: 10.1007/978-0-387-30162-4_415

24. Côté L-P, Mishara BL. Effect of helping suicidal people using text messaging: An evaluation of effects and best practices of the canadian suicide prevention services text helpline. Suicide and Life-Threatening Behavior Wiley; 2022 Aug;52(6):1140–1148. doi: 10.1111/sltb.12908

25. Gilat I, Rosenau S. Volunteers perspective of effective interactions with helpline callers: Qualitative study. British Journal of Guidance & Counselling Informa UK Limited; 2011 Aug;39(4):325–337. doi: 10.1080/03069885.2011.567327

Supplementary Files

Figures

Histogram of logits for the Hierarchical Weighted model.