Received: 13 September 2023 | Revised: 29 January 2024 | Accepted: 30 January 2024

DOI: 10.1111/sltb.13056

Suicide and Life-Threatening BEHAVIOR

ORIGINAL ARTICLE

Predicting suicide risk in real-time crisis hotline chats integrating machine learning with psychological factors: Exploring the black box

Meytal Grimland PhD[1,2] | **Joy Benatov PhD**[3] | **Hadas Yeshayahu MA**[3] | **Daniel Izmaylov MA**[4] | **Avi Segal PhD**[4] | **Kobi Gal PhD**[4,5] | **Yossi Levi-Belz PhD**[1,6]

[1]Lior Tsfaty Center for Suicide and Mental Pain Studies, Ruppin Academic Center, Emek Hefer, Israel

[2]Shalvata Mental Health Center, Hod Hasharon, Israel

[3]Department of Special Education, University of Haifa, Haifa, Israel

[4]Department of Software and Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[5]School of Informatics, University of Edinburgh, Edinburgh, UK

[6]Department of Behavioral Sciences, Ruppin Academic Center, Emek Hefer, Israel

Correspondence
Yossi Levi-Belz, Department of Behavioral Sciences, Ruppin Academic Center, Emek Hefer, 40250 Israel.
Email: yossil@ruppin.ac.il

Funding information
Israel Science Foundation, Grant/Award Number: 1302/01

Abstract

Background: This study addresses the suicide risk predicting challenge by exploring the predictive ability of machine learning (ML) models integrated with theory-driven psychological risk factors in real-time crisis hotline chats. More importantly, we aimed to understand the specific theory-driven factors contributing to the ML prediction of suicide risk.

Method: The dataset consisted of 17,654 crisis hotline chat sessions classified dichotomously as suicidal or not. We created a suicide risk factors-based lexicon (SRF), which encompasses language representations of key risk factors derived from the main suicide theories. The ML model (Suicide Risk-Bert; SR-BERT) was trained using natural language processing techniques incorporating the SRF lexicon.

Results: The results showed that SR-BERT outperformed the other models. Logistic regression analysis identified several theory-driven risk factors significantly associated with suicide risk, the prominent ones were hopelessness, history of suicide, self-harm, and thwarted belongingness.

Limitations: The lexicon is limited in its ability to fully encompass all theoretical concepts related to suicide risk, nor to all the language expressions of each concept. The classification of chats was determined by trained but non-professionals in metal health.

Conclusion: This study highlights the potential of how ML models combined with theory-driven knowledge can improve suicide risk prediction. Our study underscores the importance of hopelessness and thwarted belongingness in suicide risk and thus their role in suicide prevention and intervention.

KEYWORDS

crisis chat hotlines, hopelessness, machine learning, natural language processing, suicide

INTRODUCTION

Suicide is a complex public health problem of global importance. It takes a staggering toll on global health issues, with approximately 800,000 people dying from suicide worldwide each year (WHO, 2019). Beyond that, it is estimated that more than 25 million individuals attempt suicide worldwide each year (Turecki & Brent, 2016).

Theories of suicide prevention have struggled in their efforts to understand what factors may facilitate suicide risk (Turecki & Brent, 2016). A large body of research has shown suicide risk to be influenced by an interaction of biological, psychological, cultural, environmental, and social factors. Indeed, many factors have been suggested as related to suicide risk, such as depression, prior history of suicide attempts, and personal traits such as perfectionism, hopelessness, and loneliness (Bloch-Elkouby et al., 2020; Gvion et al., 2015; Klonsky & May, 2015; Mann, 2002; NIMH, 2020; O'Connor & Nock, 2014; Turecki et al., 2019; Van Orden et al., 2010).

A meta-analysis by Franklin et al. (2017), spanning five decades of suicide research, found that predictions of suicide thoughts and behaviors were somewhat better than chance and held some clinical utility. In recent years and based on technological advancement, researchers have begun studying the predictive ability of machine learning (ML; McHugh & Large, 2020) regarding suicide risk. One should bear in mind, however, that applying machine learning methods to enhance predictions do not always result in significantly improved outcomes compared to simpler models (Salganik et al., 2020). A recent meta-analysis (Kusuma et al., 2022) evaluated the performance of ML models in predicting suicidal ideation, suicide attempts, and death by suicide. The results were highly satisfying in terms of area under the curve (AUC) (0.84–0.88) and in specificity 0.87 (95% CI [0.84, 0.90]) with moderate results in terms of sensitivity, 0.66 (95% CI [0.60, 0.72]). Among these studies, almost half were electronic health care based, whereas others used data from social media. Only a handful of studies used natural language processing (NLP) to predict suicide outcomes (Kusuma et al., 2022; Xu et al., 2022). Notably, while these studies provided evidence that ML models can offer vital information to differentiate suicidal from nonsuicidal individuals, analyzing electronic health records and social media apps seemed to provide only a partial depiction of the individual's mental state in real-time suicidal situations, as one-sided texts cannot reveal situated conversation patterns characterized by temporal dynamics and diverse vocabularies.

Moreover, most ML studies have not incorporated psychological theory-driven knowledge to assess suicide risk, thus maintaining the gap between theory and practice. To our knowledge, only Xu et al. (2021) combined a ML model with domain knowledge in an online counseling site. Their domain knowledge-aware risk assessment (KARA) model outperformed the standard NLP model. However, they used a preliminary lexicon, which included, for example, negative emotions ("useless" and "disturbed") and interpersonal relationships ("breakup" and "reject") but failed to include many other major theory-driven factors. Much more information is needed to understand the specific psychological factors predictive of suicide risk in online crisis hotlines. These hotlines have been recognized as having a critical role in the care chain for individuals at risk of suicide (Joiner et al., 2007) due to their 24/7 access to paraprofessional support, guidance, and acute interventions (Gilat & Shahar, 2007).

To narrow this knowledge gap, this study aims to shed light on the specific psychological factors contributing to real-time predictions of suicide risk using ML techniques. Thus we aimed to open the "black box" of ML prediction to understand not only the levels of prediction but also the specific factors contributing to AI predictions of suicide risk across crisis hotline chats. To reach this ambitious goal, we examined a lexicon of linguistic representations for all the main suicide risk factors drawn from the prominent theories in the field (e.g., Van Orden et al., 2010). Then, we trained and validated a ML model on a large database of crisis hotline chats of one of the main crisis hotlines in Israel ("Sahar"). Operationally, we used AI-based NLP methods to ascertain to what extent the examined factors could distinguish suicide-related chats from nonsuicide-related chats. To our knowledge, no study has yet to explore such a broad range of risk factors for suicide ideation in real-time crisis calls using AI models. Such information would contribute to suicide theory and facilitate the design of clinical suicide prevention steps, with the key innovation being that this knowledge could help detect real-time suicide risk rather than determine suicide risk from self-report measures that may be retrospectively biased and prone to social desirability.

METHODS

Dataset

The study dataset comprised chat sessions at Sahar, an internet-based crisis hotline chat service for distressed individuals in Israel. The database comprised 17,654 chat sessions conducted over 4 years (2017–2021), which were recorded, documented, and manually labeled by the service's trained volunteers to identify the key topic raised in each session. The hotline sessions lasted up to approximately 40 min. The caller stated his age range and sex upon entry to the chat.

Outcome measure: Suicide risk versus nonsuicidal distress

Suicide risk was defined as any chat session whose caller exhibited suicidal ideation; suicide ideation refers to thoughts of ending one's life actively (wish to kill oneself; Liu et al., 2020). Of 17,564 labeled sessions, the trained volunteers of SAHAR classified 3097 chat sessions as including expressions of suicide risk (17%). To validate the volunteers' labeling, three clinical psychologists (three of the authors), specialists in suicide prevention, double-checked 600 randomly extracted chat sessions (200 classified as nonsuicidal and 400 as indicating suicide risk) and classified them blindly to the dichotomous categories of suicide risk (yes/no). The analysis yielded a satisfactory indicator of Cohen's kappa of 0.731 indicating adequate agreement between raters. Differences in tagging may be due to differences in the stimulus: The volunteer tagged the chat live, perhaps depicting the vibe more accurately than the clinical psychologists who tagged the chat retrospectively and offline.

Explanatory variables: Psychological factors-based lexicon (SRF)

Creating a psychological factors-based lexicon (SRF) included three steps. First, language representations of the main psychological factors and features relating to suicide risk were generated based on the main theory-driven factors currently recognized in the suicide literature (e.g., hopelessness, loneliness, and impulsivity). Second, we used validated questionnaires that tapped the theoretical constructs of the chosen main theories (e.g., PHQ9 for depressive symptoms, see examples in Appendix 1). Third, 200 random chat sessions labeled suicide risk were scanned to identify more language representations reflective of theoretical factors to enrich the SRF lexicon. The lexicon was also verified by suicide prevention experts to ensure that it tapped most of the important psychological factors and their correct language representations.

The SRF lexicon included 20 key categories of varied length, with the shortest category containing 20 words/phrases and the longest containing more than 400 words/phrases. Importantly, each sentence in each category of the SRF lexicon was tagged as either an increased-risk factor or a decreased-risk (protective) factor. For example, thwarted belongingness was represented either as "I feel like I don't belong in my surrounding" (increased) or "I feel people care for me" (decreased). Factors were included in the lexicon only if they were mentioned at least 500 times in the SAHAR chat dataset to focus on the more common language representations.

The language model

We used a natural language process model to detect suicide risk. We named this model Suicide Risk–Bert (SR-BERT), based on DialogBERT, a hierarchical transformer language model with performance in a wide range of discourse-related applications (Gu et al., 2021). The architecture is comprised of two parts: a transformer-based layer that encodes messages and an additional transformer layer that captures conversation structure, named Context Encoder Transformer.

We incorporated the SRF lexicon as a pretraining task. As described in a previous study (Izmaylov et al., 2023), we first chose a 5-dimension representation that outperformed the 20-dimension representation on the validation set, leading us to use this representation in the subsequent pre-training phase. In the second step, the self-supervised knowledge task (SSK task) was applied as a new pretraining task for predicting Sahar conversations in the SRF representation space. In addition to the SSK task, we implemented the three pretraining tasks defined by DialogBERT (Gu et al., 2021) for capturing several aspects of the conversation structure: message-level semantics, the conversation structure, and underlying dialogue sequential order.

Empirical methodology

We randomly split the labeled Sahar dataset into three sets: train (70%), validate (15%), and test (15%). These datasets were used throughout experiments described in detail in Bialer et al. (2022) and Izmaylov et al. (2023). To evaluate the model performance, we used ROC-AUC, widely employed in suicide detection research (Bernert et al., 2020). Additionally, we report the F2-score (Sokolova et al., 2006) for predicting the positive SR label. We compare SR-BERT with SSK with the following four baseline models:

SR-BERT without. SSK: This model omits the SSK pretraining task from SR-BERT with SSK. Apart from the SSK pretraining task, this model is identical to SR-BERT w. SSK.

Ensemble SI-BERT (Bialer et al., 2022): This is a nonhierarchical Hebrew language model representing state of the art for SR detection. It was trained on the same dataset from the Sahar organization. We reimplemented this model with the code and parameters provided by the authors and ran it on the dataset provided for this research.

SRF-based lexicon + XGBoost: An XGBoost (Chen & Guestrin, 2016): classifier based on encoding conversations over the Suicide Risk Factor (SRF) lexicon. We noted that XGBoost outperformed Random Forest and Logistic

Regression as the classifier for this baseline (and for the following two baselines).

Doc2Vec + XGBoost: An XGBoost classifier based on encoding each conversation to a 300-dimensional space using the Doc2Vec representation (Le & Mikolov, 2014).

Statistical analysis

We conducted a logistic regression analysis to compare the differences between the associations of suicide-related content chats versus nonsuicide-related chats. Adjustments were made for sex and age, and associations were reported as odds ratios (ORs) with 95% confidence intervals (95% CIs). All statistical analyses were performed using statsmodels, a Python package.

RESULTS

The prediction of machine learning model on suicide risk

We first compared the differences between BERT and non-BERT models in predicting suicide risk. Table 1 presents a comparison of the SR-BERT model with other models. As seen in the table, both SR-BERT-based models (with and without SSK pretraining) outperformed the Ensemble SI-BERT model in terms of recall, F1, F2, and ROC-AUC metrics. The most notable improvement was in the recall metric. Moreover, the additional SSK pretraining improved the SR-BERT without SSK results for all metrics except the ROC-AUC score, which remained stable. Ensemble SI-BERT achieved the highest precision, which yielded slightly better outcomes than SR-BERT with SSK. Ensemble SI-BERT exhibited a substantially lower recall score, which corresponds to lower F1 and F2 values. Notably, the BERT-based models outperformed the non-BERT models on all tested metrics.

We used the McNemar paired test for labeling disagreements (Gillick and Cox, 1989) to compare SR-BERT with SSK with the two models, SR-BERT without SSK and Ensemble SI-BERT. Statistical significance with $p < 0.05$ was demonstrated for SR-BERT with SSK versus SR-BERT without SSK and for SR-BERT with SSK versus Ensemble SI-BERT. Overall, SR-BERT with SSK substantially improved recall and F2 (17.9%) compared with BERT (13.9%), with only a slight decrease in precision performance. These results highlight the importance of psychological factors in predicting suicide risk in the ML model.

The specific predictive power of theory-driven factors to suicide risk

Following the results regarding improving prediction by psychological factors, we examined which specific lexicon factors facilitated a better prediction of suicide risk. Table 2 presents the results of the logistic regression analysis, showing the associations between the theory-driven factors and suicide risk. Hopelessness was significantly associated with the highest odds of suicide risk (odds ratio [OR] = 2.07, 95% CI = 1.61–2.38). History of suicide attempts (OR = 1.717, 95% CI = 1.603–1.839) and deliberate self-harm (OR = 1.44, 95% CI = 1.37–1.52) were also significantly associated with suicide risk as well as thwarted belongingness (OR = 1.34, 95% CI = 1.208–1.50). Depressive symptoms, although significant, were barely associated with higher levels of suicide risk (OR = 1.06, 95% CI = 1.037–1.08).

Surprisingly, bullying was negatively associated with suicide risk, as language representations of being traditionally bullied or cyberbullied were associated with relatively lower odds of suicide risk (OR = 0.81, 95% CI = 0.68–0.94). Similarly, identifying as LGBTQ was significantly associated with lower odds of suicide risk (OR = 0.78, 95% CI = 0.63–0.97). Notably, other known risk factors for suicide risk, such

TABLE 1 Comparing suicide risk prediction machine-learning models in crisis hotline chats ($n = 17{,}654$).

	Recall [%]	Precision [%]	ROC-AUC [%]	F2 [%]	Fl [%]
Doc2Vec[a] + XGBoost[b]	31.3	69.2	64.7	35.l	43.1
SRF lexicon[c] + XGBoost	55.l	67.2	76.5	57.l	60.0
Ensemble SI-BERT[d]	60.4	**70.9**	91.3	62.3	65.3
SR-BERT without SSK[e]	72.9	68.4	**92. l**	71.9	70.6
SR-BERT with SSK	**78.3**	68.9	**92.1**	**76.2**	**73.3**

[a]Doc2Vec. is a Natural Language Processing tool.

[b]XGBoost is a machine-learning system for tree boosting.

[c]SRF lexicon: Suicide risk factor-based lexicon.

[d]SI-BERT: suicidal ideation BERT.

[e]SR-BERT without SSK: suicide risk BERT without self-supervised knowledge.

as psychopathology, perfectionism, loneliness, and truancy, were not significantly associated with suicide risk, suggesting that many of these callers, with those known risks, contacted the hotline out of distress but not a suicidal one.

Contribution to suicide risk along the time of chat session

Based on the regression results, we also sought to discern the trajectories of the associations between the theory-based factors and suicide risk as the chat proceeded. Thus, we examined the associations between these theory-based factors and suicide risk as the chats' timelines progressed (recorded as a percentage of the chat). As seen in Figure 1, each factor's significant contribution to suicide risk was largely maintained throughout the chat session. More specifically, hopelessness had the highest odds associated with suicide risk at the beginning of the chat (20% of the chat, OR = 2.73, 95% CI = 2.14–8.56), continuing to be highly predictive of suicide risk during the remainder. Language representation of past suicidal attempts was the second highest factor to predict suicide risk at the beginning of the chat (20%, OR = 2.53, 95% CI = 2.1778–8.82) and continued to be predictive throughout the chat session; this was also the case with the other significant theory-based factors mentioned in Table 1. Interestingly, chats of callers identifying as LGBTQ were found to be associated with significantly lower suicide risk only at 60% of the chat session onward (OR = 0.781 95% CI = 0.616–0.990). Likewise, bullying was associated with significantly lower suicide risk only at the final part of the chat (80%, OR = 0.83, 95% CI = 0.71–0.98).

TABLE 2 Univariate analysis of the associations between theory-driven factors and suicide risk in hotline crisis chat sessions ($n = 17{,}654$).

Feature language representations	Odds ratio	*p* Value	95% CI
Hopelessness	**2.076**	2.04E-25	1.611–2.383
History of suicide attempts	**1.717**	1.39E-53	1.603–1.839
Deliberate self-harm	**1.445**	1.15E-41	1.370–1.524
Thwarted belongingness	**1.345**	6.06E-08	1.208–1.498
Depressive symptoms	**1.059**	7.99E-08	1.037–1.0826
Bullying	**0.805**	0.008	0.685–0.947
LGBTQ	**0.785**	0.025	0.636–0.970
Perfectionism	0.841	0.0947	0.686–1.030
Loneliness	1.120	0.153	0.958–1.310
Psychopathology	0.970	0.294	0.918–1.026
Truancy	0.944	0.457	0.813–1.097

Note: Bold number represents statistically significant results.

DISCUSSION

This study's primary aim was to shed light on the predictive value of psychological factors to suicide risk in real-time conversations using ML models. We aimed to explore the "black box" of the ML models and to understand the specific contribution of each ML model factor to suicide risk prediction in crisis hotline chats. To our knowledge, this is one of the very first studies to apply ML models to an online crisis dataset using a lexicon of theory-driven psychological factors.

Our findings highlight the high predictive power of a model combining ML and psychological factors regarding suicide risk in real-time chats at crisis hotlines. While some studies have shown that ML models can predict

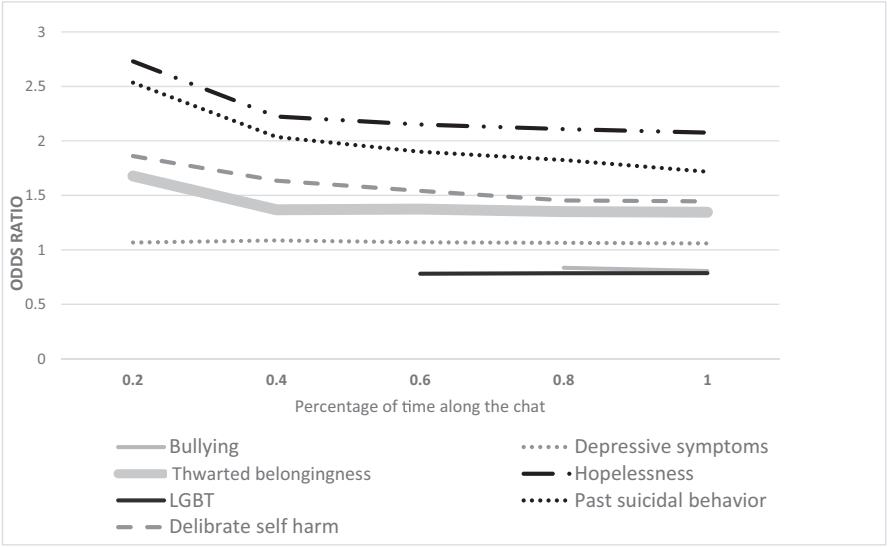

FIGURE 1 Prediction of suicide risk by theory-based factors across the time percentage of the chats ($n = 17{,}654$).

suicide risk, the current results underscore the importance of psychological factors in predicting suicide risk, above and beyond ML prediction alone. These findings support the value of incorporating theory-driven features to augment NLP-based ML models to facilitate a more effective prediction of suicide risk.

Importantly, the theory-based lexicon constructed for this study enabled us to understand the specific contribution of several psychological factors to predicting suicide risk in real-time chats at crisis hotlines. Beyond the recognized importance of previous suicide attempts and deliberate self-harm as indicators of increased suicide risk, we were able to shed light on two main theory-driven factors found as highly important for predicting suicide risk: hopelessness and thwarted belongingness. Hopelessness was found to be the most predictive factor of suicide risk in the context of crisis helpline callers (OR = 2.07). Past research has already identified it as a key factor that facilitates mental pain to progress to the level of suicide ideation and behavior (Beck et al., 1975; Galynker et al., 2014; Levi-Belz et al., 2014). It has been suggested that the negative and inflexible cognitive style characterizing higher levels of hopelessness directly affects suicide risk (e.g., Sueki, 2022). However, the current findings underscored it as the most powerful predictor of suicide risk among callers seeking help from crisis helplines, above and beyond a broad swath of psychological factors (even more predictive than a history of suicidal behavior). Interestingly, in light of our findings regarding the mild effect size of depression prediction (OR = 1.059), hopelessness can be viewed as the critical component elevating the risk among those suffering from depressive symptoms. Overall, these results highlight the central role of hopelessness in turning mental pain into suicide risk and, thus, becomes a critical factor in interventions for individuals with suicide risk.

Thwarted belongingness is defined as the experience of feeling alienated from others, particularly alluding to the painful feeling of being outside the family, friends, and other valued groups (Van Orden et al., 2010), was also found to have an important role in increasing suicide risk in our ML prediction model. The interpersonal theory of suicide, one of the leading theories of understanding suicide risk (Joiner et al., 2007; Van Orden et al., 2010), has identified thwarted belongingness as a facilitator of suicide ideation and higher suicide risk in general (Glenn et al., 2022) and in times of international crisis such as the Covid-19 pandemic (Gratz et al., 2020). Our results show that language representations of thwarted belongingness, such as "I am alone" or "There is no one I can turn to in times of need", are highly important to understanding who is at risk of suicide in real-time conversations. These results align with several studies highlighting the role of interpersonal factors in suicide risk (Klonsky & May, 2015; Levi-Belz et al., 2019).

Some studies have shown that integrating hopelessness and thwarted belongingness is critical to predicting suicide risk. Kleiman et al. (2014), examining 508 participants from a large university, found that combining both measures was the most effective predictor of suicide risk, with thwarted belongingness as a partial mediator between hopelessness and suicide ideation. Levi-Belz et al. (2014) found that the interaction of hopelessness and interpersonal difficulties, such as thwarted belongingness, was highly significant in predicting more severe suicide attempts (see also Gvion & Levi-Belz, 2018). However, whereas these studies used self-report measures of only few factors in retrospective studies, our findings underscore the predictive value of these factors in real-time suicide risk situations and with higher ecological validity.

Examining the theory-driven factors' predictive trajectories at different time points in the chats indicated that the hierarchy of the predictive risk factors was preserved from the beginning to the end of the chats. These findings further highlight the centrality of hopelessness and thwarted belongingness as key markers for detecting suicide risk at the early stages of a chat, even beyond the predictive nature of previous suicide behavior. This finding may be particularly promising for the quick detection of suicide risk at the very beginning of the chat session. In contrast to most studies, bullying and LGBTQ factors were found to be negatively associated with suicide risk. These studies found that individuals who are victims of bullying (e.g., Holt et al., 2015) or tied to the LGBTQ community (e.g., Yıldız, 2018) present an elevated risk for suicide ideation and behavior.

More research is needed to better understand the current findings. However, callers expressing language representation of bullying or LGBTQ in their chat sessions may have sought help with their difficulties and, thus, did not articulate suicide risk in presenting their issues. This observation may strengthen the notion that most of those involved in bullying or the LGBTQ community members are not at suicide risk.

The study results should be interpreted in light of several methodological limitations. First, the SRF lexicon is limited in its ability to fully encompass all theoretical concepts related to suicide risk. Moreover, using language representations of theoretical concepts (e.g., hopelessness) may suffer from several drawbacks due to its limited ability to encompass all possible language expressions of each concept. We sought to overcome this limitation by basing the lexicon on the standardized self-reports typically used to operationally examine such factors and by using suicide prevention experts to enrich it with relevant laguage expressions. However, more studies are needed to validate the relationships between each theoretical-based factor and its language

representations. Second, the dichotomous (yes/no for suicide risk) classification of chats was determined by Sahar's volunteers, who are nonprofessionals in mental health and may not have been accurate in their judgment. To overcome this important limitation, we conducted a double-crossed examination of 600 chat sessions by experts in suicide prevention who blindly relabeled them, yielding high inter-judge reliability. However, future studies should examine more objective tools for classifying chats regarding suicide risk levels. Last, as the population of callers to the Sahar online crisis helpline may not represent all populations with suicide risk and as the SRF lexicon and the chats are Hebrew-based, the study results may be limited only to the Israeli crisis helpline caller population. Future studies examining other populations could corroborate and validate this study's findings.

Conclusions and implications

The current findings highlight the critical role of ML algorithms based on NLP in predicting suicide risk in the context of real-time chats at crisis hotlines. In light of our findings regarding the centrality of hopelessness and thwarted belongingness as suicide risk indicators, these factors should be considered key elements in the early assessment of suicide risk in live crisis hotline chats. Aligning with these findings, helpers' responses should specifically target these elements and use elements such as a hope kit that includes for example recommended coping, relaxation, and distraction activities, as well as personal content that serve as reminders for positive life experiences (Vesco et al., 2022) in their interventions with callers who experience high mental pain and suicidal ideation.

FUNDING INFORMATION

This work was supported by a grant from the Israel Sciences Foundation (ISF) under Grant no. 1302/01.

CONFLICT OF INTEREST STATEMENT

None.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

ETHICS STATEMENT

The study received approval from the ethical review of the Ruppin Academic Center review board.

REFERENCES

Beck, A. T., Kovacs, M., & Weissman, A. (1975). Hopelessness and suicidal behavior: An overview. *JAMA*, *234*(11), 1146–1149. https://doi.org/10.1001/jama.1975.03260240050026

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., & Abnousi, F. (2020). Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International journal of environmental research and public health*, *17*(16), 5929.

Bialer, A., Izmaylov, D., Segal, A., Tsur, O., Levi-Belz, Y., & Gal, K. (2022). *Detecting suicide risk in online counseling services: A study in a low-resource language*. arXiv preprint arXiv:2209.04830 https://doi.org/10.48550/arXiv.2209.04830

Bloch-Elkouby, S., Gorman, B., Lloveras, L., Wilkerson, T., Schuck, A., Barzilay, S., Calati, R., Schnur, D., & Galynker, I. (2020). How do distal and proximal risk factors combine to predict suicidal ideation and behaviors? A prospective study of the narrative crisis model of suicide. *Journal of Affective Disorders*, *277*, 914–926. https://doi.org/10.1016/j.jad.2020.08.088

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (pp. 785–794) https://doi.org/10.1145/2939672.2939785

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., Musacchio, K. M., Jaroszewski, A. C., Chang, B. P., & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*(2), 187–232. https://doi.org/10.1037/bul0000084

Galynker, I., Yaseen, Z. S., & Briggs, J. (2014). Assessing risk for imminent suicide. *Psychiatric Annals*, *44*(9), 431–436. https://doi.org/10.3928/00485713-20140908-07

Gilat, I., & Shahar, G. (2007). Emotional first aid for a suicide crisis: Comparison between telephonic hotline and internet. *Psychiatry: Interpersonal and Biological Processes*, *70*(1), 12–18. https://doi.org/10.1521/psyc.2007.70.1.12

Gillick, L., & Cox, S. J. (1989). Some statistical issues in the comparison of speech recognition algorithms. In *International Conference on Acoustics, Speech, and Signal Processing*, (pp. 532-535). IEEE.

Glenn, C. R., Kleiman, E. M., Kandlur, R., Esposito, E. C., & Liu, R. T. (2022). Thwarted belongingness mediates interpersonal stress and suicidal thoughts: An intensive longitudinal study with high-risk adolescents. *Journal of Clinical Child and Adolescent Psychology*, *51*(3), 295–311. https://doi.org/10.1080/15374416.2021.1969654

Gratz, K. L., Tull, M. T., Richmond, J. R., Edmonds, K. A., Scamaldo, K. M., & Rose, J. P. (2020). Thwarted belongingness and perceived burdensomeness explain the associations of COVID-19 social and economic consequences to suicide risk. *Suicide & Life-Threatening Behavior*, *50*(6), 1140–1148. https://doi.org/10.1111/sltb.12654

Gu, X., Yoo, K. M., & Ha, J. W. (2021). DialogBERT: Discourse-aware response generation via learning to recover and rank utterances. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), 12911–12919. https://doi.org/10.1609/aaai.v35i14.17527

Gvion, Y., Horesh, N., Levi-Belz, Y., & Apter, A. (2015). A proposed model of the development of suicidal ideations. *Comprehensive Psychiatry*, *56*, 93–102. https://doi.org/10.1016/j.comppsych.2014.09.019

Gvion, Y., & Levi-Belz, Y. (2018). Serious suicide attempts: Systematic review of psychological risk factors. *Frontiers in Psychiatry*, *9*, 5656. https://doi.org/10.3389/fpsyt.2018.00056

Holt, M. K., Vivolo-Kantor, A. M., Polanin, J. R., Holland, K. M., DeGue, S., Matjasko, J. L., Wolfe, M., & Reid, G. (2015). Bullying and suicidal ideation and behaviors: A meta-analysis. *Pediatrics*, *135*(2), e496–e509. https://doi.org/10.1542/peds.2014-1864

Izmaylov, D., Segal, A., Gal, K., Grimland, M., & Levi-Belz, Y. (2023). Combining psychological theory with language models for suicide risk detection. In A. Vlachos & I. Augenstein (Eds.), *Findings of the association for computational linguistics: EACL* (Vol. *2023*, pp. 2385–2393). Association For Computational Linguistics.

Joiner, T., Kalafat, J., Draper, J., Stokes, H., Knudson, M., Berman, A. L., & McKeon, R. (2007). Establishing standards for the assessment of suicide risk among callers to the National Suicide Prevention Lifeline. *Suicide & Life-Threatening Behavior*, *37*(3), 353–365. https://doi.org/10.1521/suli.2007.37.3.353

Kleiman, E. M., Law, K. C., & Anestis, M. D. (2014). Do theories of suicide play well together? Integrating components of the hopelessness and interpersonal psychological theories of suicide. *Comprehensive Psychiatry*, *55*(3), 431–438. https://doi.org/10.1016/j.comppsych.2013.10.015

Klonsky, E. D., & May, A. M. (2015). The three-step theory (3ST): A new theory of suicide rooted in the "ideation-to-action" framework. *International Journal of Cognitive Therapy*, *8*(2), 114–129. https://doi.org/10.1521/ijct.2015.8.2.114

Kusuma, K., Larsen, M., Quiroz, J. C., Gillies, M., Burnett, A., Qian, J., & Torok, M. (2022). The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: A meta-analysis and systematic review. *Journal of Psychiatric Research*, *155*, 579–588. https://doi.org/10.1016/j.jpsychires.2022.09.050

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning. PMLR*, *32*(2), 1188–1196.

Levi-Belz, Y., Gvion, Y., Horesh, N., Fischel, T., Treves, I., Or, E., Stein-Reisner, O., Weiser, M., David, H. S., & Apter, A. (2014). Mental pain, communication difficulties, and medically serious suicide attempts: A case-control study. *Archives of Suicide Research*, *18*(1), 74–87. https://doi.org/10.1080/13811118.2013.809041

Levi-Belz, Y., Gvion, Y., Levi, U., & Apter, A. (2019). Beyond the mental pain: A case-control study on the contribution of schizoid personality disorder symptoms to medically serious suicide attempts. *Comprehensive Psychiatry*, *90*, 102–109. https://doi.org/10.1016/j.comppsych.2019.02.005

Liu, R. T., Bettis, A. H., & Burke, T. A. (2020). Characterizing the phenomenology of passive suicidal ideation: A systematic review and meta-analysis of its prevalence, psychiatric comorbidity, correlates, and comparisons with active suicidal ideation. *Psychological Medicine*, *50*, 367–383. https://doi.org/10.1017/S003329171900391

Mann, J. J. (2002). A current perspective of suicide and attempted suicide. *Annals of Internal Medicine*, *136*(4), 302–311. https://doi.org/10.7326/0003-4819-136-4-200202190-00010

McHugh, C. M., & Large, M. M. (2020). Can machine-learning methods really help predict suicide? *Current Opinion in Psychiatry*, *33*(4), 369–374.

NIMH (National Institute of Mental Health). (2020). Suicide prevention. https://www.NIMH.nih.gov/health/topics/suicide-prevention/index.shtml#part_153180

O'Connor, R. C., & Nock, M. K. (2014). The psychology of suicidal behaviour. *The Lancet Psychiatry*, *1*(1), 73–85.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. In *AI 2006: Advances in artificial intelligence, 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, Proceedings 19* (pp. 1015–1021). Springer.

Sueki, H. (2022). Relationship between Beck hopelessness scale and suicidal ideation: A short-term longitudinal study. *Death Studies*, *46*(2), 467–472. https://doi.org/10.1080/07481187.2020.1740833

Turecki, G., & Brent, D. A. (2016). Suicide and suicidal behaviour. *The Lancet*, *387*(10024), 1227–1239. https://doi.org/10.1016/S0140-6736(15)00234-2

Turecki, G., Brent, D. A., Gunnell, D., O'Connor, R. C., Oquendo, M. A., Pirkis, J., & Stanley, B. H. (2019). Suicide and suicide risk. *Nature Reviews Disease Primers*, *5*(1), 74. https://doi.org/10.1038/s41572-019-0121-0

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner, T. E., Jr. (2010). The interpersonal theory of suicide. *Psychological Review*, *117*(2), 575–600. https://doi.org/10.1037/a0018697

Vesco, K. M., LaCroix, J. M., Bond, A., Fox, A., Ribeiro, S., Darmour, C., & Ghahramanlou-Holloway, M. (2022). Three clinical techniques from cognitive behavior therapy for suicide prevention. *Social Work in Mental Health*, *20*(6), 672–681. https://doi.org/10.1080/15332985.2022.2050878

WHO (World Health Organization). (2019). *Suicide in the world*. Global Health Estimates.

Xu, Z., Chan, C. S., Zhang, Q., Xu, Y., He, L., Cheung, F., Tsang, C., Liu, J., & Yip, P. S. (2022). Network-based prediction of the disclosure of ideation about self-harm and suicide in online counseling sessions. *Communication & Medicine*, *2*(1), 156. https://doi.org/10.1038/s43856-022-00222-4

Xu, Z., Xu, Y., Cheung, F., Cheng, M., Lung, D., Law, Y. W., Chiang, B., Zhang, Q., & Yip, P. S. (2021). Detecting suicide risk using knowledge-aware natural language processing and counseling service data. *Social Science & Medicine*, *283*, 114176114176. https://doi.org/10.1016/j.socscimed.2021.114176

Yıldız, E. (2018). Suicide in sexual minority populations: A systematic review of evidence-based studies. *Archives of Psychiatric Nursing*, *32*(4), 650–659. https://doi.org/10.1016/j.apnu.2018.03.003

How to cite this article: Grimland, M., Benatov, J., Yeshayahu, H., Izmaylov, D., Segal, A., Gal, K., & Levi-Belz, Y. (2024). Predicting suicide risk in real-time crisis hotline chats integrating machine learning with psychological factors: Exploring the black box. *Suicide and Life-Threatening Behavior*, *54*, 416–424. https://doi.org/10.1111/sltb.13056

APPENDIX 1

Examples of language representations of the main theoretical concepts (translated from Hebrew).

Perceived burdensomeness	Thwarted belongingness	LGBT	Drugs and alcohol	Bullying	Depressive symptoms	Hopelessness	Loneliness	Deliberate self-harm	Past suicidal history
Better off without me	No one cares for me	Gay	Drunk	Bully me	Sad	Future is dark	No friends	Cut myself	kill myself
I let down	Outsider	Lesbian	Smoke everyday	Harassing	Not enjoying anything	No expectations	I'm lonely	Hurt myself	Commit suicide
I ruin everything	Disconnected	Transgender	Weed	No one talks to me	Depressed	I'm desperate	Isolated	Cuts are burning	Took pills
I'm damaged	Not wanted	Bisexual	I was wasted	Spread rumors	In bed all day	I'm hopeless	Always alone	I burnt myself	Gastric lavage
I'm always causing problems for others	I have no one	Queer	Drugs	Make fun of me	Can't concentrate	No hope	On my own	Cuts in my hands	Bridge to jump
I'm dragging others down	I don't belong in this world	Came out of the closet	Drink alcohol	Call me names	Lost interest in everything	Nothing to live for	I have nobody	Cuts in my legs	Drown myself
My death is worth	I'm alone			Laugh at me	I blame myself	No reason to live	No one to turn too	Self-harming	Suicide act
I'm a burden	I feel lonely			Ignore me		Nothing can change		Urge to self-harm	Suicide behavior
I hate myself				Shaming		Tired of life			